REVIEW OF METHODS USED FOR THE ESTIMATION OF FLOODS IN THE SANTA CLARA COUNTY, CALIFORNIA

A Report to the

Santa Clara County Flood Control and
Water District, San Jose, California

Prepared by

V. Yevjevich
and
D. B. Simons
Hydraulic Consultants

Fort Collins, Colorado 80521

September 1967

MICROFILMED

PG 2 OF 78

TABLE OF CONTENTS

Chapter		Page
I	Introduction	1
	1. Subject	1
	2. Objectives	1
	3. Three approaches for the estimation of floods of small watersheds from the viewpoint of type of available data	2
	4. Selection of 100-year return period flood for design purposes	4
	5. Time durations selected for description of flood volume frequencies	4
II	Data on Precipitation and Runoff for the Estimation of Floods in Santa Clara County	6
	1. Data in various reports	6
	2. Data supplied by the District	8
	3. Use of the above data	8
III	Use of Precipitation Data Only For Estimation of Floods in Santa Clara County	9
	1. Approach in the use of precipitation data	9
	2. Error in rainfall measurement	9
	3. Error in estimation of total rainfall for a storm over a drainage basin	10
	4. Errors in estimation of parameters in any rainfall-runoff model	11
	5. Total errors	12
	6. Final remarks	12
IV	Simultaneous Use of Rainfall and Runoff Data for Flood Frequency Analysis in Santa Clara County	14
	1. Partial use of precipitation data	14
	2. Simultaneous use of rainfall and runoff data	14
	3. Conclusions	16
V	Runoff Data Available for Estimation of Flood Characteristics	17
	1. General availability or runoff data	17
	2. U.S. Geological Survey stations	18
	3. Data of the Santa Clara Valley Water Conservation District	19
	4. Data from reservoir level observations	20
	5. Mixing of data of various accuracies	21
VI	Probability Functions Best Suited for Flood Frequencies	22
	1. Functions for flood frequencies	22
	2. Selection of flood frequency function	23
	3. Selection of flood frequency function for the analysis of floods in the Santa Clara County	24

MICROFILMED

PG 3 OF 13

TABLE OF CONTENTS (cont)

Chapter		Page
VII	Confidence Limits	27
	1. Sampling variations of frequency curves	27
	2. Two ways to determine confidence limits for log-normal distribution of flood characteristics	33
	3. Confidence limits of the frequency curve, represented as regression straight line	37
	4. Use of confidence limits	39
	5. Basic recommendation for the District studies	40
VIII	Estimate of Parameters for Flood Frequency Curves	42
	1. Procedure used by the District	42
	2. Estimation of central value parameter of flood frequency curves for various watersheds.	44
	3. Estimation of dispersion parameter of flood frequency curves for various watersheds.	51
	4. Remarks on the use of step-wise method in selecting the order of significant independent variables.	57
	5. Potential new independent variables.	58
	6. Skewness coefficient of frequency curves	58
	7. Significant independent variables for prediction equations	59
	8. Effective number of independent stations	60
IX	Regional Distribution of Regression Constants (Isopleths).	62
	1. District's procedure	62
	2. General criticism of use of regression constants.	64
	3. Use of regression constants in the Districts work	65
X	Flood Hydrographs	67
	1. District's procedure	67
	2. Discussion of balanced hydrograph	67
	3. Other possible approaches	68
XI	Effects of Urbanization	71
	1. The continuous process of urbanization	71
	2. District's procedure to determine the effect of urbanization on floods	72
	3. Comments on the above District's procedure	73
REFERENCES		77

MICROFILMED

PG 4 OF 5

CHAPTER I

INTRODUCTION

1. Subject

The Santa Clara County Flood Control and Water District, San Jose, California, hereinafter referred to as the District, has developed methodology for the estimation of flood characteristics on its watersheds. This methodology is based chiefly on runoff data. A review of procedures of this methodology, with suggestions for its improvement is the subject of this report. Computations and values in this report should be used only for selection of methodology or as the basis for an interim flood frequency report, since all data in this report are used only to further the appraisal of methods and procedures or to show the merit of improvement suggestions.

2. Objectives

The basic objectives of this report are:

(a) An analysis of various methods and procedures used in the estimation of flood frequencies of streams in Santa Clara County, with an assessment of positive and negative aspects of the methodology presently used by the District.

(b) Suggestions for an improvement in methodology, to be used in the forthcoming interim hydrology report. Suggestions are such that information on flood characteristics of various watersheds may be obtained for planning and design work by the District, and that the flood protection for the County may be continued with the most reliable data and methods currently available.

(c) Suggestions for a long range program of improving the methods to be used in the estimation of flood characteristics, assuming a

MICROFILMED
12-1-87

better location of gaging station, and the collection of proper data, and for a general improvement in accuracy of flood estimation.

(d) Evaluation, in general terms, of the character and reliability of the available hydrologic data from the point of view of its use to estimate the flood characteristics in Santa Clara County with the best possible accuracy.

3. Three approaches for the estimation of floods of small watersheds from the viewpoint of type of available data

The estimation of flood characteristics of a small watershed or a group of small watersheds may be based on precipitation data only, on runoff data only, or jointly on precipitation and runoff data. Let's assume that there is an amount of information, I_p, about floods in the given quantity of available precipitation data, and an amount of information, I_r, about floods in the given quantity of available runoff data. The high degree of correlation between the simultaneous data of runoff and precipitation makes the information amounts, I_p and I_r, two mutually dependent factors, because large parts of their content are repetitious. In principle, the joint use of the simultaneous precipitation and runoff data may produce an amount of information, I_{pr}, which may be somewhat larger than either I_p or I_r. It will be shown later that the District uses both types of data, though precipitation is used to a lesser extent.

If precipitation data have been gathered over the same period as runoff data, and if both precipitation and runoff have been observed with the same accuracy, the precipitation data will add little to the runoff data for the determination of flood characteristics. Under particular conditions, the use of precipitation data may decrease

FILMED
PG 6 OF 10

information or reliability of flood characteristic estimates. This point is not usually appreciated, especialy by those eager to use all data available to evaluate properties of floods. Because of the nature of precipitation observations,there are often substantial errors in precipitation data, and the pulling together of the two types of data--the relatively accurate runoff data and the less accurate precipitation data--may lead to a loss of information which is contained in runoff data and consequent reduction in accuracy of estimation.

The case is different if the precipitation and runoff are only partially simultaneously observed, or if the data on precipitation are over a much longer period of observation than the data on runoff. By pulling together both types of available data one would be able to obtain a maximum information I_{pr} which may be significantly greater than I_r. Therefore, the use of those methods which will extract the maximum information from a total quantity and quality of hydrologic data should be the guide line in selecting the best methodology for the estimation of properties of floods in Santa Clara County.

The general analysis of estimation methods of flood properties in the area of Santa Clara County is reviewed here from the following three points of view:

(1) Use of precipitation data only;

(2) Use of runoff data only; and

(3) Use of both precipitation and runoff data.

The term "estimation of floods" here includes: (a) determination of probability distribution (approximated or estimated by frequency curves) of flood peak discharges; (b) determination of probability distribution of flood volumes which are estimated by flood volume

MICROFILMED
12-1967
PG 2 OF 28

frequency distribution curves; (c) estimate of flood hydrographs of various return periods; and (d) the computation of confidence limits for the above flood frequency curves to allow for potential sampling errors in estimated floods. The term "estimation of floods" is used here with the implication that all curves or values determined represent only approximations to the true population curves or values, or estimates of factors which--by definition of hydrologic populations of types such as floods or droughts--are never known.

4. Selection of 100-year return period flood for design purposes

The selection of a fixed return period of floods in Santa Clara County streams seems reasonable for the present problems of flood control. This comes from the desire to offer the same degree of flood protection for all in the County. However, as investments in the flood plains grow continuously it may be advantageous to change the return period of floods as the basis for design in some densley populated areas. Therefore, the method of flood estimation, and particularly the accuracy of frequency curves should be considered in such terms as to produce floods for return periods even greater than 100 years.

5. Time durations selected for description of flood volume frequencies

The 1-day and 2-day flood volumes have been selected to describe the flood volumes for the purpose of studying the effect of general purpose storage reservoirs or flood control reservoirs of floods along the streams in Santa Clara County, or for construction of a

FILMED
12-1967
PG 8 OF 78

total hydrograph for flood routing purposes. From these 1-day and 2-day volumes a construction of a "balanced hydrograph" is made. Since drainage basins have a large range of areas, from a couple of square miles to hundreds of square miles, it is questionable whether these two variables of flood volumes describe the flood hydrographs of such a wide range of areas equally well. It would be more advantageous to make the time duration for these flood volumes a function of the drainage basin size.

Regardless of this effect of basin area on the selection of these durations, the approach made by the District in selecting the three frequency curves of (1) annual maximum peak discharge; (2) annual maximum 1-day flood volume; and (3) annual maximum 2-day flood volume, enables an investigation of the change of flood frequency parameters with the duration of flood volumes, starting from zero duration of the flood peak to the 2-day duration.

MICROFILMED
12-1967
PG 9 OF 71

CHAPTER II

DATA ON PRECIPITATION AND RUNOFF FOR THE ESTIMATION OF FLOODS IN SANTA CLARA COUNTY

1. Data in various reports

The report "Hydrologic Data Index, Santa Clara County, California" by Carroll E. Bradberry and Associates, 1963 Revision, summarizes the climatological and surface water data available in 1962, and is a revision of the 1962 inventory of data [1]*. Climatologic data include: precipitation, temperature, evaporation, humidity, wind and insolation data. Surface water data include: streamflow, river stage, lake level, tide level and sediment transport data. Data for estimates of floods within the County are taken from observations at stations both within the County and the immediately surrounding area. Data are classified according to reliability and other factors.

The report "Hydrologic Atlas, Santa Clara County, California" by Carroll E. Bradberry and Associates, 1964 Revision [2], represents a partial analysis of accumulated data, organized into charts, tables, and maps. The analysis includes synthesized data which is obtained by transforming the precipitation data into runoff data by a special Stanford Watershed Model.

The report "Synthesized Streamflow Data, Santa Clara County, California" by Carroll E. Bradberry and Associates, 1964, presents the data on runoff during the period 1924-1962, which is obtained by transforming the precipitation data into runoff data (daily flows) by the use of Stanford Watershed Model [3].

---

*References are designated by [ ] and are given at the end of this report.

ILMED
12-1967

The methodology used for the synthesis of runoff data is described in the "Summary Report, Hydrology Data Program, Phase III-Streamflow Synthesis, Santa Clara County," by Carroll E. Bradberry and Associates, 1964 [4].

The report "Santa Clara County Flood Control and Water Conservation District, Hydrology Study - Las Animas Creek" by Water Resources Engineers, Inc., of December 27, 1965, contains data which are useful for the purpose of evaluation of methodology used for the flood estimation [5] by the District.

The report "Reconnaissance Report on Floods of January 21-24, 1967" by the Santa Clara County Flood Control and Water District, January 1967 [6] gives an insight into the type of floods which occur on the streams of this County. For gaging station on six streams, this flood was estimated of having return periods of 13-17, 8, 3, 12, 15 and 11 years.

The "Report of Survey for Flood Control and Applied Purposes, San Francisquito Creek, Santa Clara and San Mateo Counties, California," by the U. S. Army Corps of Engineers, San Francisco, January 1951 [7], gives valuable information on floods of one of the County streams. It shows a large diversity in flood hydrograph shapes ranging from a sharp one-peak hydrograph to a very flat hydrograph and multi-peak flood hydrographs.

The "Soil Survey of Santa Clara Area, California," issued by U. S. Department of Agriculture and California Agricultural Experiment Station [8] gives a good insight into soil characteristics of various watersheds.

FILMED
1967
PG 11 OF 22

The "Report on Standard Project, Rain-Flood Criteria, of Sacramento-San Joaquin Valley, California," U. S. Army Corps of Engineers [9] gives the various criteria and approaches developed by the Corps for the analysis of floods.

2. Data supplied by the District

Apart from the above reports, the District has supplied the following documentation and data to the writers of this report:

(1) Short write-up by the District entitled,"Regional Hydrologic Frequency Study," which is the basic document on the methodology for flood frequency analysis and which is reviewed in this report. This report is, therefore, a general review of various procedures, techniques or methods described in that document.

(2) Annual flood series and their frequency curves for many stream gaging stations.

(3) A map of isopleths of regression constants.

(4) Several flood hydrographs of various streams.

(5) Topographic maps of the County.

(6) Topographic map of watersheds for which runoff data are used in flood frequency analysis.

3. Use of the above data

The above reports and data supplied by the District were a sufficient documentation for the purpose of this report in assessing various aspects of methodology for flood frequency analysis and in suggesting for improvements in that methodology.

MICROFILMED
12-1967
PG 12 OF 18

CHAPTER III

USE OF PRECIPITATION DATA ONLY FOR ESTIMATION OF FLOODS IN SANTA CLARA COUNTY

1. Approach in the use of precipitation data

An attempt to use precipitation data (hourly values) to obtain runoff (daily values) was made by Carroll E. Bradberry and Associates who applied the Stanford Watershed Model. Simultaneous observations of rainfall and runoff were utilized to determine the parameters of the Model. The Model was then applied to transform hourly rainfall into daily streamflows of 17 County watersheds, for which no runoff data were available. The synthesized data approach was also applied to fill gaps in runoff data on several other drainage basins.

The basic question arises, how accurate are the flood peak values or flood volumes from synthesized data in comparison with the flood peaks or flood volumes which could be observed? Similarly, one is interested in the comparison between flood hydrographs from synthesized data and actual flood hydrographs. Before any discussion of the methodology based on the Stanford Watershed Model, one must question the general reliability of precipitation data. How representative is the precipitation data available of the actual rainfall which occurred over the entire drainage basin?

The errors which are inherently present in the synthesized runoff data of the Santa Clara County are manyfold, and some of those errors may be very high. They are briefly discussed here.

2. Error in rainfall measurement

The average difference between measured rainfall and actual rainfall at a given point and in a storm which produces the large floods--say

of a return period of 2 years or greater--is likely to be around ±10% of the true value. Sometimes this difference is much greater. This is either a random error or often is a combination of random and systematic errors. Systematic errors are frequent, due to changes in gage location or elevation or changes in gage environment such as the growth of trees.

3. Error in estimation of total rainfall for a storm over a drainage basin

The difference between the total storm rainfall over a drainage basin estimated from a limited number of gaging stations and the actual fallen precipitation depends basically on these three factors:

(a) Number of stations which cover a given area (surface area per one gage);

(b) Uniformity in the distribution of gages over the drainage basin; and

(c) General variation of rainfall over the area, particularly during large storms.

All three sources of this type of error greatly affect the accuracy of determination of rainfall during storms over the drainage basins in Santa Clara County. First, the number of precipitation gaging stations of long records is small. Only two stations on the County lower floor (San Jose and Palo Alto) are long-record stations. Second, the number of stations with sufficient records (say 20-25 years) is also small. Third, the density of stations over drainage basins is not sufficient. Fourth, the distribution of stations over basins is far from being ideal. Fifth, the variation of precipitation over

MICROFILMED
12-1967
PG 14 OF 11

mountain drainage basins is large, and the ratio of maximum average annual precipitation to the minimum average annual precipitation over the same drainage basin varies from 2:1 to 4:1. This variation may be much greater from storm to storm.

The rapidly varying precipitation intensity, duration and areal coverage over both the east and west slopes of the County mountains is an important factor in assessing the value of precipitation data for synthesizing of daily runoff. It justifies the statement that any determination of rainfall amounts over a drainage basin during a storm from a small number of gaging stations must necessarily have a very limited accuracy.

4. Errors in estimation of parameters in any rainfall-runoff model

Regardless of the type of rainfall-runoff model used for transforming rainfall data into runoff data, many parameters must be estimated from the simultaneous data on rainfall and runoff. Whoever tries to determine a unit hydrograph as the measure of drainage basin response to surplus unit rainfall knows well the difficulties in estimating the losses, the portion of rainfall which is infiltrated, and the uncertainties in separation of the groundwater portion of observed hydrographs from the surface runoff contributions. Besides, the unit hydrographs obtained from individual storms of the same basin vary sometimes within large limits. Errors of various types are compounded in a relatively inaccurate rainfall-runoff model which is synthesized in the form of average or approximate unit hydrographs.

The Stanford Watershed Model may be conceived as an advanced model of river basin response. Many parameters of this model must be adjusted in such a way that the observed rainfall input produces the

observed runoff output. Theoretically, some general properties of the drainage basin response, considered as "black box," are postulated and several parameters must be estimated by a trial-and-error method so that the known inputs reproduce the observed outputs. Mathematically speaking, the Stanford Model is an indeterminate system because the relations of various phases of drainage basin response may have various shapes and various parameters and still produce a match between the observed and the synthesized runoff. Besides, the model must be adjusted and parameters must be estimated in such a way as to produce the most accurate results at the average or most frequent discharges. It is to be expected that the largest deviations between the synthesized and the observed discharges should be for floods of very small probabilities. So many imponderables are involed in any rainfall-runoff model that substantial errors are inevitable.

5. Total errors

By summing up the variances of these three basic types of errors, and by assuming that they are basically independent, the total variance is obtained. The probable error is defined as that value of the total error in which the chances of the error being either greater or smaller than the probable error are equal, 50%. One should expect the probable error of estimated floods by rainfall synthesis to be very high, reaching even 25-50%, and in many cases even higher.

6. Final remarks

The use of precipitation data for the estimation of floods in the Santa Clara County should be considered, at least at the present status of data available, as a less reliable method than the use of runoff

MICROFILMED
12-1967
PG 16 OF 78

data. The use of precipitation data could be accepted only in case the runoff data cannot produce flood characteristics of a much greater accuracy.

The use of precipitation data for the estimation of flood characteristics for the Santa Clara County watersheds may be attractive in the future when much larger pool of data becomes available and more has been learned about the distribution of rainfall in time and space over various drainage basins in the County. However, the rainfall data may be an important variable in determining the relationships between the frequency distribution parameters of flood characteristics and the drainage basin and rainfall parameters.

It is recommended that rainfall data be given a secondary role in the forthcoming interim hydrology report on flood characteristics in the Santa Clara County. Runoff data of about 40 stream gaging stations in the Santa Clara County and around it will produce much more reliable information on floods than the synthesis of precipitation data could yield. The secondary role of precipitation means only that the rainfall data with its most accurate properties (say annual or seasonable values of precipitation) should be used for the estimation of parameters of flood frequency distributions on ungaged watersheds.

FILMED
12-1967
PG 17 OF 18

CHAPTER IV

SIMULTANEOUS USE OF RAINFALL AND RUNOFF DATA FOR FLOOD FREQUENCY ANALYSIS IN SANTA CLARA COUNTY

1. Partial use of precipitation data

The approach used by the District of putting the main stress on flood frequency analysis by using the runoff data is sound. In the light of all hydrologic data available for drainage basins in Santa Clara County and surrounding area, this approach will produce the most reliable information on floods of large return periods. However, since drainage basin area is not the only factor which changes the flood frequency parameters from one basin to another, and since the amount of precipitation changes from one basin to another, it is also necessary to use precipitation data to evaluate parameters of flood frequency curves. This has been done by the District. It is a proper approach, and it represents a partial use of precipitation data.

2. Simultaneous use of rainfall and runoff data

It may be advocated that the runoff data should be used for the estimates of flood characteristics in the Santa Clara County for periods of runoff observations, but that the precipitation data should supplement runoff data for other periods for which rainfall data is available but runoff data is lacking. In this way, the total periods of observation would be increased at many runoff gaging stations. For some drainage basins the runoff can be predicted from rainfall effectively only if very short periods of simultaneous observations of rainfall and runoff are available. According to this reasoning, runoff data of very short periods (say 2-5 years) cannot be effectively used to derive frequency curves, but it can very well be used to establish rainfall-runoff

MICROFILMED
12-1967
PG 18 OF 78

relationships for that basin. Then, applying this derived relationship, the long-term observations of rainfall can be transformed into long series of runoff data.

This theoretical approach is very attractive and can be effectively used under some conditions. However, many dangers may be encountered if this approach is not critically evaluated in each specific case. For flood estimates in drainage basins of Santa Clara County, this approach needs a very careful evaluation. Flood estimation from rainfall by the best methods of rainfall-runoff relationships will have much smaller accuracy in the County when compared with estimates of flood characteristics from runoff gaging stations operated by the U. S. Geological Survey. It is quite likely that the accuracy of floods determined from rainfall, especially under the prevalent condition of high variability of rainfall with altitude in this County, is also much less than the estimates of floods from the runoff data of gaging stations operated by the Santa Clara Valley Water Conservation District. One would be inclined to rank floods determined from rainfall as much less accurate than the flood volumes estimated from reservoir levels of San Francisco Water Department.

One may well question the watered-down accuracy of flood estimate information obtained from a mixture of relatively accurate runoff data and highly inaccurate rainfall data. The answer to this question is not simple, but in the case of the Santa Clara County the writers would suggest a cautious attitude in an eventual mixing of the above two types of flood estimate information. When better and more numerous data on rainfall become available, one would be

FILMED
1967
PG 14 OF 22

logically tempted to prove that the use of rainfall data in addition to runoff data in the above manner may increase information on floods.

It is recommended to the District that the rainfall data should not be used to synthesize flood events for the forthcoming interim hydrology report. However, it is recommended that a 5-year improvement program and the future revision of the interim report should be planned in such a way that proper rainfall data may be collected and then used for synthesizing flood events to supplement flood estimates from runoff data. This should be the case particularly for the watersheds with ungaged streams, or with gages of runoff observations for short periods of time.

3. Conclusions

In summary, the use of the rainfall-runoff relationship and the use of rainfall data for flood estimates should be omitted for the interim hydrology report of the District. However, a development of relationships between the flood frequency parameters and the rainfall data by regression analysis is necessary for the interim report. The rainfall data were used in the District methodology in the form of average annual precipitation. For the type of precipitation regime in Santa Clara County with high seasonal rainfall, this means that the average precipitation during the rainfall season is used for the purpose of predicting flood frequency parameters. The daily rainfall or rainfall of smaller time intervals or other time intervals such as storm duration, have not been used by the District for the estimation of flood frequencies. In the opinion of the writers of this report, that approach by the District was correct.

CHAPTER V

RUNOFF DATA AVAILABLE FOR ESTIMATION OF FLOOD CHARACTERISTICS

1. General availability or runoff data

Runoff data is available both within and from areas surrounding the County. Data from outside the County are used to supplement the scarce long-term runoff data in the County, to enable a more reliable determination of isopleths of some parameters at the County's boundary and to generally increase the accuracy of estimates of flood characteristics. Runoff gaging stations with more than nine years of observation were used in the study. They were favorably distributed across the area of interest. The watersheds of these stations range in size from 3 to 1,200 square miles, in a total area of about 4,000 square miles. The large majority of stations was located on watersheds in the size range of 5 to 50 square miles. From the data, 28 stations were used for the analysis of flood peaks, and all 39 stations available were used for the analysis of maximum one-day and two-day flood volumes.

The time series of runoff data in the Santa Clara County flood study are composed of three types of data of essentially different quality or reliability:

1. Data from the stream gaging stations operated and observed by the U. S. Geological Survey. This type of data is most reliable for the estimation of flood peaks and flood volumes.

2. Data from the stream gaging stations operated by the Santa Clara Valley Water Conservation District. These data are very reliable for low and median discharges but have poor accuracy for flood peak and flood volume estimation. A recent revision of the upper part (flood part) of rating curves improved somewhat the accuracy of estimated floods.

MICROFILMED
12-1967
PG 21 OF 1

3. Data from reservoir inflows, computed by determining the changes in water elevations of City of San Francisco reservoirs and of the Santa Clara Valley Water Conservation District reservoirs, during flood periods. These data were used only for one-day and two-day flood volume estimations. They make the difference between 39 total stations used and 28 stations used only for flood peak estimation.

The above data were scrutinized for various possible inconsistencies, effects of flow regulation, diversions and other irregularities. Some adjustments were made for certain annual events when justified by the District investigators.

Many inaccuracies are inherent in the data, due mainly to errors produced by gaging instruments and by data processing methods. The principal errors probably come from the extrapolation of rating curves to flood stages, sometimes for several orders of discharge magnitude beyond those for which the discharge measurements were made. It was considered by the District investigators that detailed historical studies of individual station records were not feasible when dealing with so large a sample of stations. This attitude toward a detailed review of historical background and quality of data of 39 stations should be seriously reconsidered, because the results on flood characteristics cannot be any better than the accuracy and general reliability of the basic data. The improvement in accuracy of data by the recent revision of rating curve for stations of the Water Conservation District is an illustrative example.

2. U.S. Geological Survey stations

The data from the stations operated by the U.S. Geological Survey should be considered as the basic runoff information for the estimation of both flood peaks and flood volumes. The reasons are:

12-1967
PG 22 OF 13

(a) Stations are operated to produce the most reliable results for any flow (low, medium, high) within the limits of funds available and the feasibility of measurement of extremely large flows;

(b) Standard methods of gage installation, checking, discharge measurement and rating curve determination are used for all stations in approximately the same manner;

(c) Continuous checks of gaging stations and revisions of rating curves lead to a persistent improvement of runoff data reliability.

It is a generally accepted fact that the estimation of flood discharges is of a much lower accuracy than for the discharges around the median flow. The value of relative probable error (a ratio of 2/3 of the standard error in estimated discharge to the true discharge) is often mentioned to be about 10%. In other words, by using the data of U. S. Geological Survey stations the estimates of peak discharge have an error of about 10%, although this varies from station to station and depends on how many discharge measurements have been made close to the estimated flood peak.

Data on flood characteristics from U. S. Geological Survey stations can be considered as the primary quality station data, provided that any change in flow regulation, in water diversion and in the other man-made changes can be taken into account by the proper corrections and computations. The basic question which arises in the use of this primary data is whether or not data of lesser reliability should be mixed with this primary quality data.

3. Data of the Santa Clara Valley Water Conservation District

Runoff observations made by the Santa Clara Valley Water Conservation District at several stations is of a lower accuracy than the U. S. G. S.

for flood discharges. The main errors result from the extrapolation of measured discharges to high stage values to produce the rating curve for flood flows without measurements close to or in the range of flood discharges.

It is likely that some flood peaks estimated from this type of data may be in error by 20-30% or more. They are of a different order of accuracy in comparison with the U. S. Geological Survey data. The mixing of data of two different orders of accuracy is bound to decrease the general accuracy of results. Therefore, the data on floods from the Water Conservation District should be considered as secondary quality station data. It should be very carefully used in the derivation of basic conclusions and final design flood characteristics, regardless of recent revisions.

4. Data from reservoir level observations

Of the three types, these data are least reliable. The measurement of the change of levels in a reservoir and derivation of the basic characteristics of flood inflow hydrographs therefrom is always considered in any hydrologic service to be of a low accuracy. It is known that the determination of the first derivative from cumulative hydrologic curves is subject to substantial errors, even if the cumulative curves are carefully observed. However, when these curves are subject to substantial errors, the accuracy of determining the first derivative (say flood peak discharge in the case of reservoir level observations) or the volume increments (one-day maximum inflow, or two-day maximum inflow) is not satisfactory. It is sufficient for the blowing wind to change intensity and direction during the period of observation to substantially affect the accuracy of the observed data.

MED

The longer the period for which the inflow volume is considered the better is the accuracy of the derived information. The use of data from reservoir level observations is much more accurate for two-day flood volumes than for one-day volumes, and even more accurate than for the flood peak discharge.

Data from reservoir level observations should be considered of the third rank importance in comparison with the previous two types of data. The Santa Clara County Flood Control and Water District was right in using this data only for the one-day and two-day maximum flood volumes. However, even for such flood volumes this data on floods should be considered of a lesser reliability than the data derived from the U. S. Geological Survey data.

5. Mixing of data of various accuracies

One must ask the basic question, whether the mixing of these three types of data may improve the general accuracy for flood frequency parameters, or whether that mixing may decrease the overall accuracy. This is a crucial question which cannot be answered directly. However, it is significant that there are steps in accuracy between these three types of data. It is, therefore, attractive to consider three levels of data processing namely, (1) only primary stations*are used; (2) primary and secondary stations combined; and (3) all three types of station data are pulled together in one ensemble.

*The definition of primary and secondary stations in this report is different from the meanings given to these two terms in the District's methodology of flood frequency analysis

CHAPTER VI

PROBABILITY FUNCTIONS BEST SUITED FOR FLOOD FREQUENCIES

1. Functions for flood frequencies

The District used both the Pearson Type III probability distribution function as applied to logarithms of flood peak discharge, 1-day and 2-day maximum flood volumes, and log-normal probability distribution function. The basic idea was the use of a unique value of skewness coefficient for all gaging stations for Pearson Type III function.

In general, three basic probability distribution functions* have been used by hydrologists to fit to cumulative frequency curves of observed flood series. They are:

(1) The log-normal probability density function, or normal probability density function applied to logarithms of flood events, or

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma_n\,x} e^{-(\log x - \mu_n)^2/2\sigma_n^2} \qquad (1)$$

where x = flood event, μ_n = mean of logarithms of x, σ_n = standard deviation of logarithms of x, and f(x) = probability density of x. The third parameter is sometimes used in the form of lower boundary x_o; in this case, x on the right side of Eq. (1) is replaced by $(x - x_o)$.

(2) The extreme largest value probability distribution function, also called the double exponential function, or the Gumbel probability distribution function

$$F(x) = \exp\left[-e^{-\alpha(x - \beta)}\right] \qquad (2)$$

with β = mode, and $F(\beta) = e^{-1} = 0.38$; α = the scale parameter = $1.281/\sigma$, where σ is standard deviation of x, and F(x) is the probability of floods smaller than x.

*The following three probability functions, eqs. (1) through (3), are given also in other forms in the literature.

(3) Three parameter Gamma probability density functions, equivalent to Pearson Type III function

$$f(x) = \frac{1}{\beta\,\Gamma(\gamma)} \left(\frac{x - m}{\beta}\right)^{\gamma-1} \exp\left[-\left(\frac{x - m}{\beta}\right)\right] \qquad (3)$$

where m = lower boundary, β = scale parameter, γ = shape parameter, and $f(x)$ = probability density of the flood event x.

The above parameters, μ_n and σ_n in Eq. (1), α and β in Eq. (2), and m, β and γ in Eq. (3) are related to the three basic parameters of any distribution, μ (mean), σ (standard deviation) and C_s (coefficient of skewness).

Sometimes, x in Eq. (3) is replaced by $\log x$, a departure from the Gamma types of distribution curves.

2. Selection of flood frequency function

The extrapolation of flood frequency to large flood return periods can be made in two ways:

(a) By a graphical extrapolation of plotted frequency curves. If a straight line fit is feasible, one extrapolates that line; and

(b) By fitting probability function to frequency curves, and estimating in the appropriate way the function parameters. From that function, one can obtain flood values for any probability (or return period).

This second approach gives a better accuracy if the estimate of parameters is carried out by the most reliable estimation procedure. The combination of a graphical and analytical approach is usually the most feasible. The plot of frequency curves in the appropriately transformed coordinates shows whether a function fits the data well. The estimate of parameters by the best estimation method gives a probability distribution equation from which the flood of a given return period may easily be determined.

MICROFILMED
10-1-07

Only the double exponential probability distribution function has a theoretical justification for its use on flood peak discharges. On the Gumbel paper, the frequency curve of the flood peak discharge should follow well a straight line. It was, however, shown that in many cases the log-normal probability distribution function fits as well not only for the frequency curves of flood peak discharge, but also and particularly for the frequency curves of flood volumes. Practice has shown that the Pearson Type III also fits the same frequency curves well.

One is tempted to look at these three potential probability distribution functions with a purely theoretical consideration. In this case, stress should be given to the Gumbel double exponential function. However, it has not yet been proven that it is applicable to flood volumes but only to flood peaks. If one looks from the practical side and especially in the light of the high sampling variations of frequency curves for small samples and from the viewpoint that large errors exist in the computed flood values, one is tempted to use the log-normal distribution for both the peaks and the flood volumes. The Pearson Type III distribution may be applied either to peaks or flood volumes.

U. S. Geological Survey publication "Flood-Frequency Analysis," Manual of Hydrology, Part 3, Flood-Flow Techniques [10], gives a detailed description of the use of the double exponential function for flood frequency curves.

3. Selection of flood frequency function for the analysis of floods in the Santa Clara County

The use of Pearson Type III function has several disadvantages:

FILMED
PG 28 OF 13

(1) It uses logarithms of flood events instead of events themselves. That part of the expression in Eq. (3)

$$\exp\left[-\left(\frac{\log x - m}{\beta}\right)\right] \tag{4}$$

has two cancelling algebraic operations, and substantially changes the distribution curves.

(2) It is difficult to find coordinate scales which are transforms of f(x) and x or log x such that the probability distribution curves (or cumulative frequency curves) become straight lines, though attempts have been made to develop them (Alexander, Australia).

(3) It is more difficult to carry out the statistical inferences by determining confidence limits than is the case for log-normal distribution.

In the light of the quality and quantity of flood flow data of the 39 runoff stations used in the District's work on floods, the simplest probability function to use is the log-normal distribution. The advantages of its use are as follows:

(1) It can be applied equally to flood peak discharges and to 1- and 2-day flood volumes;

(2) Distribution curves plot as straight lines on paper with log-probability scales;

(3) It is very simple to determine the confidence limits for any confidence level;

(4) Extrapolation beyond the observed values does not produce values that deviate substantially from observed values; and

(5) There is no need to use a constant skewness coefficient because each slope of the curve (or standard deviation of logarithms)

determines the skewness coefficient uniquely. The skewness coefficient depends on the third statistical moment of frequency curves, and the estimation of this moment is highly unreliable. Therefore, it is difficult to test whether the skewness coefficient of flood frequency curves is constant, and the estimation of this moment is highly unreliable. Therefore, it is difficult to know whether the skewness coefficient of flood frequency curves is a constant or not.

When using the Pearson Type III probability function or some other distributions derived from it, hydrologists in the U.S.S.R. often employ an approximation of $C_s = 2C_v$, inferring that the skewness coefficient is double the coefficient of variation, with $C_v = \sigma/\mu$, or the ratio of standard deviation to the mean. It has been shown by these hydrologists that this approximate ratio $C_s = 2C_v$ seemingly works well under very different conditions. It is not suggested to the District that this relationship of $C_s = 2C_v$ should be used for flood frequency analysis of the Santa Clara County streams. It is discussed here simply to question the advisability of using a constant value of C_s for all flood frequency distributions of the County in the Pearson Type III distribution.

Considering all the above points, the log-normal probability distribution is recommended to be used for frequency curves in the interim hydrology report of the District, to be issued by Fall, 1967. However, for the future 5-year program of updating this interim hydrology report--by further data collection and by revisions of various procedures and computations--the use of all three probability functions is recommended: log-normal, double exponential and Pearson Type III, and a comparison of results obtained by each of them.

MICROFILMED
12-1967
PG 30 OF 1

CHAPTER VII

CONFIDENCE LIMITS

1. Sampling variations of frequency curves

It is easy to conceive a population of flood events which has a log-normal distribution. Plotted in log-probability scales, it gives a straight line. It is also easy to simulate thousands of events from this population by the data generation method (Monte Carlo method). From these data, many samples of size n may be created.

Several factors and properties may be studied about these samples of size n, but for the purposes of this report, interest is on the mean of logarithms (value for 50% probability in log-probability scales), the standard deviation of logarithms (one-half difference between values for probabilities 84.13% and 15.87%), and whether the plotted frequency curves may be fitted by straight lines or not.

Because of simple procedure in making statistical inference with log x, where x = flood variables and log x is normally distributed, one can easily obtain the sampling variation of the mean and standard deviation of logarithms. However, the conclusion is not so simple for the test of how often the plotted lines would depart from a straight line, though the population distribution is a straight line in log-probability scales.

Benson [10], pages 51-74, has made these latter tests by using Gumbel double exponential function and Gumbel paper for straight line plots of this distribution, and 1000 simulated extreme largest numbers. He plotted 100 frequency curves, each of the sample size n = 10 (Figure 1a); 40 frequency curves, each of the sample size n = 25 (Figure 1b);

20 frequency curves, each of the sample size n = 50 (Figure 1c); and 10 frequency curves, each of the sample size n = 100 (Figure 1d).

It is sufficient to glance at these four figures to ascertain that, even for sample sizes of n = 50 and n = 100, several frequency curves are far from straight lines though their population curve is a straight line. For frequency curves of sample sizes of n = 10 and n = 25--which approximate most of the record lengths of river gaging stations used in the District's study of floods in Santa Clara County--it becomes surprising how many of them depart from straight lines in the Gumbel paper.

The conclusion is that one should expect many flood frequency curves in the District's study not to plot as straight lines in log-probability scales, though the population distribution would probably be close to a straight line and follow the log-normal distribution. The requirement and the expectation are false that all or a predominant number of 39 or 28 frequency curves, with first number for 1-day and 2-day flood volumes and second number for the flood peaks, should follow straight lines in log-probability scales in order to assume that floods are log-normally distributed.

It is sufficient that a simple majority of plotted flood frequency curves are close to straight lines in log-probability scales in order to justify the use of the log-normal distribution for flood events. This is what occurs with the frequency curves in the District's study.

Therefore, one is seriously tempted to suggest to the District a very simple approach, namely to use the log-normal function for flood frequency curves only, and to base the statistical inference on this function for the interim report.

2. Two ways to determine confidence limits for log-normal distribution of flood characteristics

Assume that the District will use the log-normal distribution for the plotting of flood data and fit straight lines to points plotted by the m/(n+1) plotting position. The confidence limits* may be determined in two ways:

(a) By putting confidence limits on the mean of logarithms and the standard deviation of logarithms of frequency curves; and

(b) By expressing the straight line distribution curves in log-probability scales in the form of linear regression lines, and determining confidence limits about these straight lines.

The first approach is simpler than the second, but both are described in the following text.

(1) The probability level of confidence limits should be first selected. This decision must be made by the District if one does not intend to follow the conventional levels for confidence limits. Explicitly, the less risk (probability) District likes to have for values to fall outside the confidence limits, the greater should be the probability level. It is suggested that the District should use a 75%-80% level (12.5%-10% probability of values being outside any confidence limit), except if there is a particular reason for a higher level to have a larger confidence interval by selecting 90%, 95%, 97.5% or 99% levels in order of running a smaller risk for values to fall outside confidence interval.

(2) If the flood data approximately follow the straight lines in graphs with log-probability scales, the logarithms of flood values approximately follow the normal probability distribution function. This fact makes the determination of confidence limits for the two parameters, the mean and the standard deviation of logarithms, much simpler than for

*The confidence limits relate to fitted probability (frequency) distribution curves

FILMED 1987

non-normal distributions. In the light of approximations inherent in the straight line fits, the accuracy in parameter confidence intervals which are determined by straight line fits is sufficient in the case of log-normal distributions.

(3) The mean and standard deviation of logarithms may be determined in two ways, either from data by proper calculation or from the straight lines fitted. The use of computed values of all flood events may create a distortion in the case of exceptionally low values of lowest floods in the sample. This is the case with several stream stations in Santa Clara County. Therefore, the straight line fit is a better approach for Santa Clara County. If the calculation approach is used and values are expressed in the form log x_i (where x_i is any value of flood variable), the mean of logarithms $y = \log x$ is

$$\mu = \frac{1}{n} \sum_{i=1}^{n} \log x_i , \tag{5}$$

and the unbiased estimate of standard deviation of logarithms is

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (\log x_i - \mu)^2} . \tag{6}$$

For the graphical estimate of the mean and standard deviation of logarithms, a sketch is included in Figure 2 to show the procedure.* For 50% probability, the mean of logarithms from the fitted straight line is obtained. The difference of logarithms $y_1 - y_2 = 2\sigma$ gives two standard deviations of logarithms, where y_1 is log x_1 for 15.87% probability and y_2 is log x_2 for 84.13% probability. The value of 2σ gives a better accuracy than the difference taken for y for the above probabilities individually from μ or from the value of 50% probability.

*For this figure a large confidence interval (95% level) is selected.

FILMED
1967

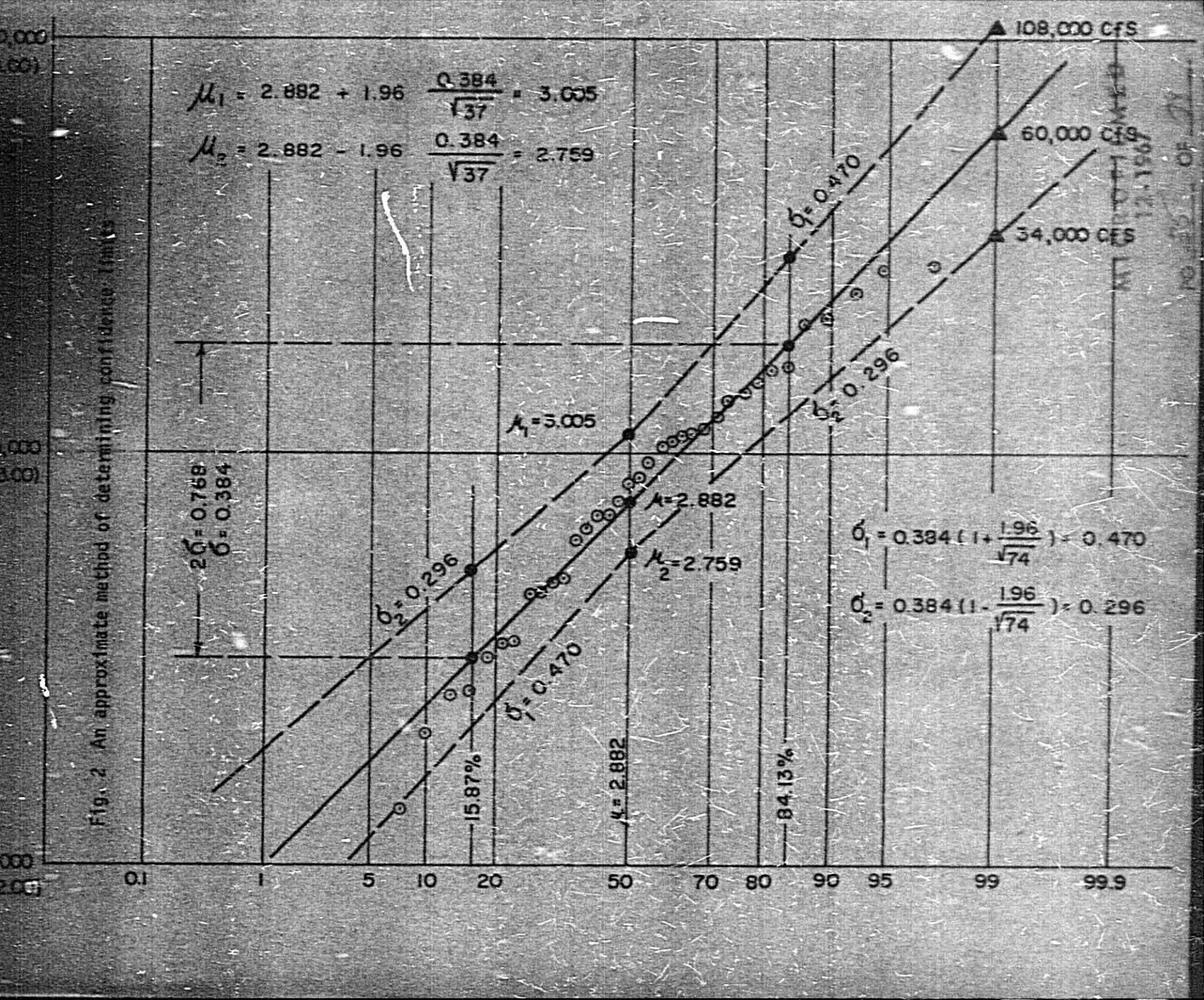

Fig. 2 An approximate method of determining confidence limits

(4) The determination of confidence limits for the mean of logarithms is made for the selected confidence level of 95%, which states that each tail of the probability distribution has 2.5%. The value t on both tails of standardized normal variable for 95% level is $t = \pm 1.96$ (for the level 90%, $t = \pm 1.65$; for the level 99%, $t = \pm 2.58$).

For the mean of logarithms, μ, and the unbiased (calculated or graphically determined) standard deviation of logarithms, σ, the values of μ_1 and μ_2 are designated as confidence limits of μ. They are determined in such a way that the probability of the true mean being between μ_1 and μ_2 is 95% (or any other selected probability level), so that

$$\mu_{1,2} = \mu \pm \frac{t\,\sigma}{\sqrt{n}} \tag{7}$$

where n is the sample size. For 95% confidence level, $t = 1.96$, so that

$$\begin{aligned} \mu_1 &= \mu + 1.96\,\frac{\sigma}{\sqrt{n}} \\ \mu_2 &= \mu - 1.96\,\frac{\sigma}{\sqrt{n}} \end{aligned} \tag{8}$$

These values are shown schematically in the accompanying graph, Figure 2, as μ_1 and μ_2.

(5) Confidence limits for the standard deviations of logarithms are determined similarly as for the mean of logarithms. The confidence limits σ_1 and σ_2 for the standard deviation of logarithms σ may be obtained by a simple procedure in the case of the normal distribution of logarithms of flood values as

$$\sigma_{1,2} = \sigma \pm \frac{t\,\sigma}{\sqrt{2\,n}} = \sigma\left(1 \pm \frac{t}{\sqrt{2\,n}}\right) \tag{9}$$

For 95% probability level, the probability is 95% that the true value of standard deviation of logarithms is between

$$\sigma_1 = \sigma \left(1 + \frac{1.96}{\sqrt{2n}}\right)$$

$$\sigma_2 = \sigma \left(1 - \frac{1.96}{\sqrt{2n}}\right) \qquad (10)^*$$

(6) Confidence limits of the frequency distribution lines are obtained by using σ_1 for the upper right portion of confidence limit though μ_1, and σ_2 for its lower left portion, and thus the upper confidence limit is approximated. Similarly, σ_2 for the upper right and σ_1 for the lower left portions of the lower limit through μ_2 gives the lower confidence limit as an approximation. These limits are shown in the accompanying graph, Figure 2. Each confidence limit line represents two straight lines which intersect at μ_1 and μ_2, respectively, in the graph paper with log-probability scales. However, for Cartesian scales, they represent two curves which intersect at values of x corresponding to μ_1 and μ_2.

From the frequency curve itself, and from the right leg of the upper confidence limit curve, the two values of 100-year return period floods are obtained. These values then become the basic flood design values of 100-year return period.

3. Confidence limits of the frequency curve, represented as regression straight line

Replacing the probability scale in the accompanying graph by a linear scale as the variable $z = m/(n+1)$, then the straight line for $y = \log x$ is

*Equations (7)-(10) are based on the assumption of normal distribution of sample means about the population value μ, and the sample standard deviations about the population value σ. However, for small n (less than 10) Student-t distribution should be used for the distribution of sample means (when μ is replaced by m), and Chi-square distribution should be used for the distribution of sample standard deviations (when σ is replaced by s).

$$y = a + b\,z \tag{11}$$

which is a regression straight line fitted to the points log x and m/(n+1). Then the confidence limits* can be computed by

$$y_{1,2} = y_i \pm t_p \hat{s}_y = \bar{y} + b\,(z_i - \bar{z}) \pm \frac{t_p \hat{s}_{ey}}{\sqrt{n}} \sqrt{1 + \frac{(z_i - \bar{z})^2}{\hat{s}_z^{\,2}}} \tag{12}$$

In this equation, y_i = log x_i is any value from the fitted straight line; $\bar{y}$ = the mean of logarithms (at 50% probability); b = the regression coefficient of the fitted straight line; z_i = a variable taken as the positive linear distance (in the accompanying graph) right of 50% probability and as the negative linear distance left of 50% probability (the selection of the linear scale for z is not important because it affects the b-coefficient), and in this case $\bar{z}$ is zero; t_p = the value from the Student t-distribution, which can be obtained from tables of this distribution as soon as the probability level is selected and the sample size n is known, with n - 2 degrees of freedom. To obtain either $\hat{s}_y$ or $\hat{s}_{ey}$, the best approach is to compute the correlation and regression coefficients r and b of y against z, for $\bar{z}$ = 0 as

$$r = \frac{1}{n\,\hat{s}_y\,\hat{s}_z} \sum_{i=1}^{n} y_i z_i \tag{13}$$

with the unbiased value of s_y and s_z as

$$\hat{s}_y = \sqrt{\frac{1}{n-1} \sum_{1}^{n} (y_i - \bar{y})^2} \quad \text{and} \quad \hat{s}_z = \sqrt{\frac{1}{n-1} \sum_{1}^{n} z_i^{\,2}} \tag{14}$$

with y_i the values for n plotted points (and not from the straight line) for given values of z_i, and

*These confidence limits are not very wide. However, the following limits may be used for errors in $e_i = y_i - y$, (where y_i = plotted value, y = value from fitted straight line, e_i = residuals) as wider limits

$$\eta_{1,2} = \bar{y} + b\,(z_i - \bar{z}) \pm t_p\,\hat{s}_{ey} \sqrt{1 + \frac{1}{n} + \frac{(z_i - \bar{z})^2}{n\,\hat{s}_z^{\,2}}}\ .$$

$\eta_{1,2}$ are confidence limits for e_i .

MIC... 12-1...

$$b = \frac{1}{n\,\hat{s}_y^{\,2}} \sum_{1}^{n} y_i\, z_i \qquad (15)$$

$$\hat{s}_{ey} = \sqrt{\frac{n(1 - r^2)s_y^{\,2}}{n - 2}} \qquad (16)$$

4. Use of confidence limits

The use of confidence limits as determined in the above section should be as follows: (1) the designer has to select the confidence level. By selecting 90% limits, the designer is assessing that there is only a 5% chance for 100-year flood to be higher than the value of the upper confidence limit and a 5% chance to be lower than the value of the lower confidence limit. In other words, the probability that the true 100-year flood taken from the frequency curve itself may be greater is only 50%, and 50% that it may be smaller. All three flood characteristics, peak discharge, 1-day and 2-day flood volumes, should be determined both from the frequency curve and from the upper confidence limit.

The question arises whether in computing flood hydrographs (and their associated volumes) one should compute a hydrograph for each possible confidence limit flood (and volume) and separately route these hydrographs, in addition to those from the frequency curves. The answer to this question is that the District should determine 100-year frequency peak discharge and 100-year 1-day and 2-day flood volumes from the frequency curves, as well as 100-year peak discharge and 100-year 1-day and 2-day flood volumes from the upper confidence limit The "balanced hydrograph" or any other design hydrograph should be determined for both, and the flood routing should be drawn also for both of these hydrographs.

The question then arises whether would it be within the accuracy of the methodology (prediction and routing) to merely apply the same

percent correction obtained as the ratio of the upper confidence limit flood and the most probable 100-year flood from the frequency curve, to a single routed hydrograph obtained from the frequency curve. The answer is no, for the simple reason that this approach would assume a linearity in the relationship between the channel volume and the channel depth. If the channel volume increases faster than depth, which is usually the case, then the routing of the hydrograph obtained from the upper confidence limit will give a relatively lower maximum level than if a simple linear relationship is applied to the level or discharge of routed hydrograph which is obtained from the frequency curve.

5. Basic recommendations for the District studies

The following is recommended:

1. That the District use graph paper with log-probability scales for the flood frequency distributions.

2. That straight lines be fitted to plotted points (either by eye fit, or by least squares method offitting straight lines) and only fitted to the upper right portion of frequency curves when the lower left portion shows anomalously low values.

3. That the mean and the standard deviations of logarithms of flood distributions be estimated either by given expressions, or graphically as shown above, keeping in mind that the graphical estimates are less accurate but are feasible for the approximate results and conditions of some flood frequency curves in Santa Clara County flood frequency analysis.

4. That the first method of determination of confidence limits for the log-normal frequency distribution curves be used by the District,

MICROFILMED
12-1967
PG 40 OF 72

by estimating the confidence limits of the mean and the standard deviation of logarithms and by then drawing the confidence limit as shown in Figure 2. The accuracy of this method corresponds closely to the accuracy of graphical fitting procedure.

5. For the use of the second method of determination of confidence limits, the straight regression lines should be obtained by the least squares method for those portions of frequency curves which are devoid of anomalously low flood values.

6. For the ungaged watersheds of Santa Clara County the values of μ and σ will be predicted by their regression prediction equations. Both μ and σ have errors in their predicted values. These errors are measured by the standard error of residuals (or standard error of regression equation). As μ and σ refer to logarithms of flood values (mean of logarithms and standard deviation of logarithms, respectively), the standard errors of regression equations, σ_μ and σ_σ also relate to logarithms of flood events, respectively for the mean (μ) and standard deviation (σ). A simple but approximate method of computing the confidence limits in this case is to assume that σ_μ and σ or σ_σ and σ of eqs. (7) through (10) are independent, so that the composite value of the two is $\sigma_C = \sqrt{\sigma_\mu^2 + \sigma^2}$ for the mean and $\sigma_C = \sqrt{\sigma_\sigma^2 + \sigma^2}$ for the standard deviation confidence limits. By replacing σ in eqs. (7) - (10) by these two composite values, first in eqs. (7) and (8) and second in eqs. (9) and (10), the confidence limits* for flood frequency curves of ungaged watersheds may be determined.

*How to determine the value of sample size for ungaged watersheds, see eqs. 26b and 26c, and the accompanying text on pages 60 and 61

CHAPTER VIII

ESTIMATE OF PARAMETERS FOR FLOOD FREQUENCY CURVES

1. Procedure used by the District

For flood frequency curves the District has adopted the logarithmic Pearson Type III distribution, defined by the mean, the standard deviation and the skewness coefficient of logarithms of the annual maximum flows. The regional skewness coefficient was used as the mean of skewness coefficients of long-term stations. Only the mean and standard deviation of logarithms were used as varying parameters.

To obtain series of annual flood values of peak discharge, 1-day and 2-day volumes, stations were sorted into geographically related groups. For each group a primary station with the longest and most reliable flow data was selected, and the remainder of the stations in the group were termed secondary stations. The gaging station on the Arroyo Secco, with 65 years of records, was considered as the base station.

The available flood data of a primary station were correlated with that from the base station, and the correlation coefficient was obtained. The parameters of flood frequency distributions of peak discharge, 1-day and 2-day volumes of primary stations were modified in relationship to parameters of the base station by equations suggested by Langbein:

$$m_1' - m_1 = (m_2' - m_2)\, r\, \frac{s_1}{s_2} \qquad (17)$$

$$s_1' - s_1 = (s_2' - s_2)\, r^2\, \frac{s_1}{s_2} \qquad (18)$$

where 1 refers to the primary station and 2 refers to the base station, the primes are long-term values determined by the length of base station record, non-primes are the corresponding values of the length of primary

station records, and m and s are the mean and standard deviation of logarithms of annual flood values, respectively.

When many stations are used to derive regional characteristics of flood frequencies, a very common practice is that the data of stations with the longest record and with the most reliable data are systematically used to fill the missing records of stations with shorter length of observation or of less reliable data. In this approach, all the data of stations with data extended or corrected by the correlation analysis or otherwise are pulled together in order to derive regional characteristics of flood frequencies. This procedure has a basic characteristic. It repeats information manyfold which is already contained in data of stations with the longest and best records. The above procedure followed by the District is correct, if one wishes to put an excessive weight on the records of the base station. However, if the Arroyo Secco should have a systematic error or a substantial sampling error in the data, these errors would automatically be preserved and propagated by this approach.

A procedure similar to the foregoing has been used for the relationship between secondary stations and the primary station.* In this way, the information from the base station has been twice passed down this scale, and all station data contains the same possibility of error. Too much weight is put on the base station rather than on the primary stations and least on the secondary stations.

This position can be defended with many valid arguments. Therefore, there is strong support for the method adopted by the District in passing information on floods from base station and primary stations down the rank order of stations as described above.

---

*Definitions of primary and secondary stations by the District differ from the same terms as defined in this report.

MICROFILMED
12-1967

Another approach may be also useful and defended with valid arguments. In using various groups of stations,* primary stations (including the base station) would constitute the first group; all primary and secondary stations would belong to the second group; and all stations, primary, secondary and third class stations, as defined in the previous text, would be part of the third group.

The first two groups would be used for frequency analysis of flood peak discharge, and the first and third groups would be used for the frequency analysis of 1-day and 2-day flood volumes. No transfer of information would be made between stations by correlation analysis, and no correction of parameters m and s would be made by Eqs. (17) and (18). Instead, the length of record of each station would be used as weight for the parameters of individual stations. This procedure would be repeated for each group, twice for the peak discharge and twice for flood volumes. The difference obtained between the groups would give an insight into whether there is a substantial difference in information between the primary stations, on one side, and all stations available for flood peaks or flood volumes, on the other side.

2. Estimation of central value parameter of flood frequency curves for various watersheds

It is a current experience in a hydrologically homogeneous region that the flood frequency curves--in the coordinate system of peak discharge (or flood volume) versus the return period (or the probability) of floods--shift upward both with an increase of the drainage basin area and with an increase of precipitation. One of several central value parameters may be selected for the purpose of defining frequency curves, such as the mean flood, the median flood, and the mean value of logarithms of floods. Each of these basic parameters determines the location of

* Primary, secondary and third class stations as defined previously in this report.

MICROFILMED 12/1967

frequency distribution curves on the graph. It is the first and the most important parameter.

The procedure used by the District to relate that parameter to geometric factors of drainage basin, and to precipitation is a sound approach. The availability and accuracy of this equation for central value parameter is one of the key factors in any reliable prediction of floods for the Santa Clara County system of streams.

The procedure used in deriving the central value parameter--in the District's study it is the mean of logarithms of annual maximum flows--was as follows: The means of logarithms of primary stations* were adjusted according to Eq. (17) in relation to the base station. Then the same equation was used to adjust the means of secondary stations in relation to the primary station. All means represented the dependent variable in the following equation:

$$\text{Log } Q = \log C + s \log A + t \log P + u \log S + v \log Sh \qquad (19)$$

or in the equation

$$Q = C\, A^{s}\, P^{t}\, S^{u}\, Sh^{v} \qquad (20)$$

where

Q = flood value corresponding to m, with m = log Q, and m = mean of logarithms;

A = watershed area;

P = mean annual precipitation;

S = slope parameter of the basin;

Sh = shape parameter of the basin,

and C, s, t, u, and v are regression coefficients.

Equation (20) is well selected and fits the experience obtained with several flood frequency studies (notably Benson's study of New England

*Primary and secondary stations as defined by the District.

floods). The logarithmic transformation of Eq. (19) represents the case of linear multiple correlation and regression analysis. Several objections to the use of this transformation may be cited, among which the following two are the most important:

(1) The equation gives more weight to smaller values of variables; and

(2) It assumes that logarithms have a small change in the variance of deviations of sample points (a point with given values of Q, A, P, S and Sh) from the fitted function of Eq. (20), regardless of the value of log Q, while the variance of these deviations increases for an increase of Q. This assumption is only partly correct.

Regardless of these objections, the District was correct in adopting the above equations and procedures for the estimation of the mean of logarithms of flood flows for various watershed basins in Santa Clara County.

The analytical estimates of μ and σ by Eqs. (5) and (6) from flood series has a basic disadvantage. In many flood series, the low flood peaks and low 1-day and 2-day flood volumes are exceptionally small. Figure 3 shows a typical example which gives the flood frequency curves for the peak, 1-day and 2-day volumes of the Coyote Creek at NR Madrone. The left sides of these curves show very low values. In Eqs. (5) and (6), they will produce smaller values of μ and larger of σ than one would obtain by using a graphical fit to the main part of the curve over the right 90% of probability. As the flood values for less than 10% and greater than 90% of probabilities have much greater sampling errors, a graphical fit of curves through points between 10% and 90% of probabilities, and even occasionally between 20% and 80% of probabilities, may give more reliable estimates of μ and σ than by use of Eqs. (5) and (6).

12-1967

PG 96 OF 78

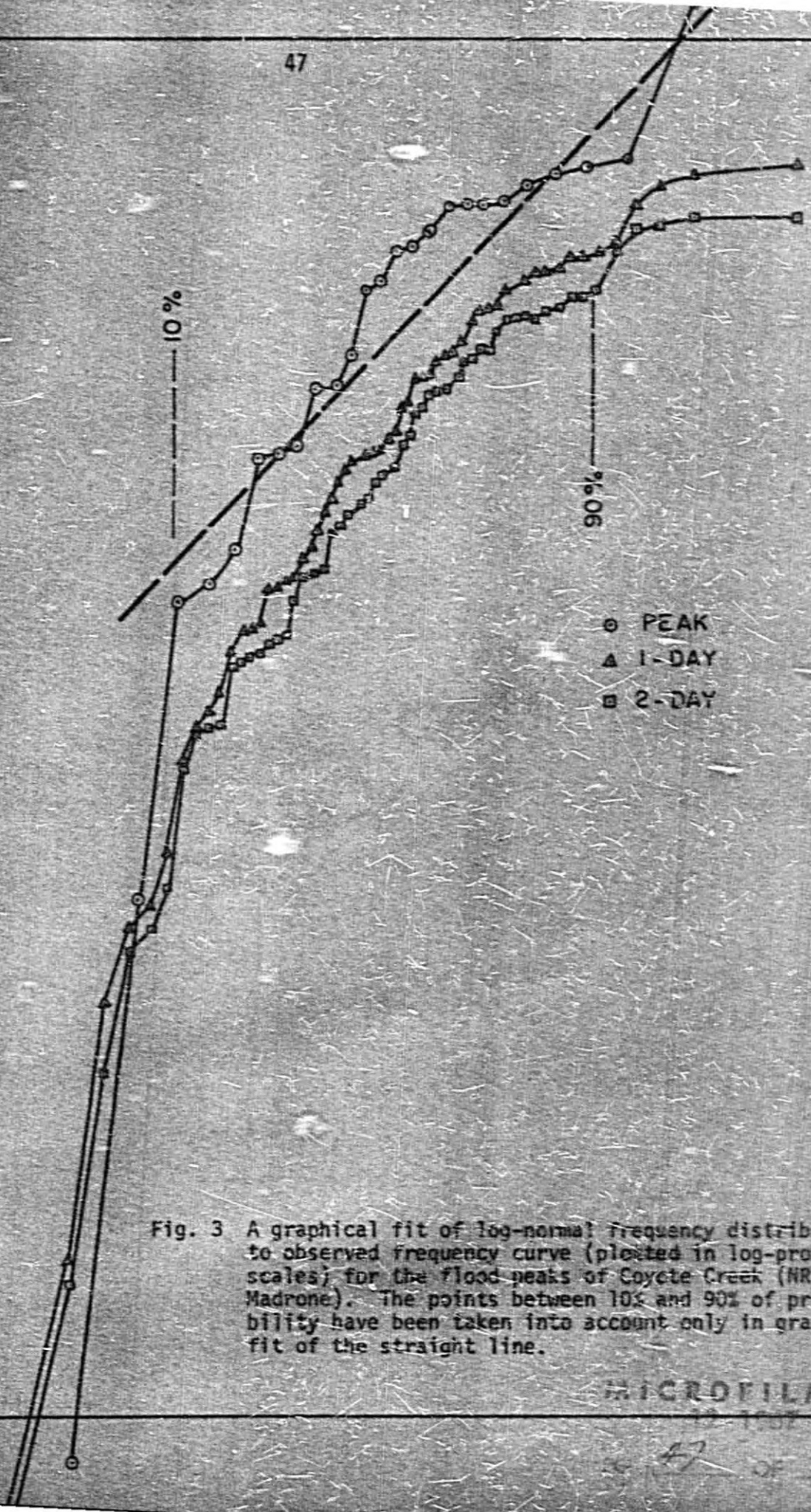

Fig. 3 A graphical fit of log-normal frequency distribution to observed frequency curve (plotted in log-probability scales) for the flood peaks of Coyote Creek (NR. Madrone). The points between 10% and 90% of probability have been taken into account only in graphical fit of the straight line.

MICROFILMED

The above Eq. (20) for the determination of Q (with m = log Q) for all three flood variables: peak discharge, 1-day and 2-day volumes, expressed in cfs or in acre-feet, respectively, has produced in the District's study the following three equations:

$$Q_{peak} = 0.411\ A^{0.763}\ P^{1.433} \tag{21}$$

$$Q_{1\text{-}day} = 0.017\ A^{1.03}\ P^{1.89} \tag{22}$$

$$Q_{2\text{-}day} = 0.013\ A^{1.037}\ P^{1.891} \tag{23}$$

with the following multiple correlation coefficients of regression analysis by Eq. (19): 0.907, 0.933 and 0.931, respectively. These coefficients are sufficiently high to give a feeling of their reliability. Any multiple linear regression equation of the type of Eq. (19) conveniently explains the largest part of variation in m, or in Q in the above Eqs. (21) through (23). The partial correlation coefficient of Q against A, for P kept constant, would probably show a very high value. Two factors are responsible for the fact that neither S (slope) nor Sh (basin shape factor) have not been shown to be significant variables in Eq. (19):

(a) Slope S and shape factor Sh are highly related to area A by geomorphologic laws of drainage basins. Since the area is the basic independent variable in Eq. (19), the step-wise regression method of selecting the sequence of significant variables in multiple linear regression analysis logically singled out first the area A. This selection then automatically accounted for the variation in the other two independent variables, S and Sh. This can be shown by computing the simple correlation coefficients of S against A and Sh against A for the watersheds analyzed. These two coefficients are likely to be high,

MICROFILMED
12-1987
PG 48 OF 78

and that would explain why the area A has accounted also for the variation of S and Sh.

(b) Area is a very dominant independent variable, so that eventual and potential small remaining contributions by S and Sh in explaining the variation of Q, beyond their variation which was already accounted for by the area A, were not considered or shown to be significant.

An alternative approach may be to exclude the large effect of the area from the basic data and Eq. (19) or Eq. (20) and to relate the remaining variation of flood characteristics to various other parameters than those included in Eq. (20).

For this approach, the dependent variable would be the unit runoff yield in floods, measured either in cfs per square mile for flood peaks, or in acre-feet per square mile, for flood volumes, as

$$q = \frac{Q}{A} . \tag{24}$$

This transformation would automatically reduce the importance of the area, A. As Eqs. (22) and (23) show for the Q-values of 1-day and 2-day volumes, the area is included with a power close to unity.* In other words, the Q values are nearly proportional to the area A. The flood peak value of Q, Eq. (21), shows a power of approximately 0.75. Therefore, by using a unit flood yield of Eq. (24), $q_{1\text{-day}}$ and $q_{2\text{-day}}$ will be nearly independent of area, and q_{peak} will be related to area by an expression of approximately $1/A^{0.25}$. The consequencies of this transformation from Q to q would be:

1. Multiple regression coefficient will drop significantly in the case of only two independent variables, A and P;

2. Precipitation would become the basic independent variable which would account for most of the variation in q;

*In a later run of revised data in the District, these powers of A in eqs. (22) and (23) were reduced to about 0.86.

3. Other watershed geometric variables, such as slope, shape, and stream density may become significant independent variables and may contribute to the increase of R-value for the dependent variable q;

4. Other watershed and precipitation variables, such as antecedent moisture conditions,* rainfall parameters other than the average annual precipitation, soils, climatic variables and others can contribute significantly to the explained variance of q.

5. The use of q as a dependent variable will undoubtedly result in a smaller value of R than those shown above for Q of the order of 0.90 - 0.93. This expectation is logical because flood characteristics depend heavily on the watershed area, and its removal as the predominant independent variable must necessarily reduce the R-coefficient. This does not mean that the prediction of q with an equation of the type of Eq. (20) with a smaller R-coefficient will be any less accurate than the prediction of Q by Eq. (20) with a larger R-value.

6. Because the prediction of Q by Eqs. (21) through (23) has only about 13% - 17.5% of unexplained variance, the use of regression constants C_p and C_2,--as it is described in the District's basic write-up on methodology,--for the regionalization approach by isopleths may seem questionable. Would it be logical to expect that these small percentages of unexplained variance of Q would be only attributable to diversity of watershed factors which vary from basin to basin and not included into the regression equations, instead of being attributed to errors in data and to unavoidable and large sampling errors? The use of a prediction equation for q-values, with much smaller explained variance of q, would more readily justify an eventual use of isopleths for the regression constants.

MICROFILMED
1967
OF

*For the future revisions of interim hydrology report, the antecedent moisture conditions should be introduced. For predictions on ungaged basins the use of joint probability of precipitation and antecedent moisture conditions will be necessary to use.

It is recommended to the District that, with the final development of Eqs. (21) through (23) for Q-variable, similar equations should be developed for q-variable, with an attempt to include in these new equations several new (so-called) independent variables of multiple linear regression.

3. Estimation of dispersion parameter of flood frequency curves for various watersheds

If flood frequency curves in their cumulative form (flood distribution curves) can be approximated by straight lines in the appropriate transformation of coordinates, the slope of the straight line becomes an important parameter. Two questions arise with this approach: (1) when is one entitled to consider that a straight line fit is a good approximation to plotted points; and (2) what should be the best coordinate transformation in order to obtain these straight line approximate fits. The Chi-square test could be used for a given level of significance to assess whether the straight line fit is acceptable. The decision about the best mathematical probability distribution for flood frequency curves determines the transformation required for the flood frequency coordinates. This problem is discussed under the chapter on mathematical probability functions for flood frequency distributions. Although the time has passed to initiate the use of the Chi-square test in the forthcoming flood frequency report, this approach should be used for the future program of increasing information on floods in the County and updating the interim report.*

A logical question is whether the slope of frequency curves significantly changes from watershed to watershed, or can it be considered as a constant for all County streams? The consideration of this problem is

*In the application of Chi-square test the extremely low values of floods should not be used. Also, the probability level of significance should not be too large.

12-1967
PG 51 OF 5

given particular attention in this subchapter, and hinges on whether or not there is sufficient information to prove that the slopes of frequency curves are significantly different from a constant. The methodology used by the District is based on the assumption that the slope is a variable parameter. It is dependent on the watershed average annual precipitation.* Its regression constant should be regionalized in the County in a manner similar to Q, considered as the central value parameter. The sequence in the analysis of the slope of frequency curves (in the case of the use of log-normal probability distribution for flood frequency curves this slope is measured by the standard deviation of logarithms of flood events) should be

1. Test whether the slopes of frequency curves, or s of Eq. (6), are or are not significantly different from a constant;

2. In case this slope is shown to be significantly different from a constant, the prediction regression equations should be developed for flood peak, 1-day and 2-day flood volumes;

3. Decide whether the use of isopleths for the regression constant is meaningful; and if so,

4. Develop a map of isopleths.

As was discussed in a study by M. A. Benson [10] entitled "Characteristics of Frequency Curves Based on a Theoretical 1,000-Year Record," it isshown that for hundred frequency curves of 10-year return periods the spread of these curves is very large (Fig. 1a) even though the data come from the same population. Similarly, 40 frequency curves of 25-year return periods (Fig. 1b) show large fluctuations of slope though a smaller variation than for 10-year return periods, and similarly, the variation of slope

*The larger the annual precipitation, the larger should be the storms which generate floods, and the losses and infiltration become less important.

has been shown to decrease substantially for 20 curves of 50-year return periods, and even more for 10 curves of 100-year return periods. Figures 1a, 1b, 1c and 1d taken from that report show clearly how greatly the slopes vary for the same population when the observation period of samples is decreased.

Instead of using this Monte Carlo procedure in simulating a 1,000-year period, a simple approach is adopted in this report to show similar results. It will be assumed that a flood probability distribution of the population is a straight line in log-probability scales. With this assumption, the logarithms of flood peaks or flood volumes are normally distributed. In this case the standard deviation of logarithms of flood peaks and flood volumes is a measure of the slope and is normally distributed.*

For a number of runoff stations used for this test from Santa Clara County and the adjacent area,--shown in Table 1--the standard deviations of logarithms of flood peaks are given as measures of slope. Table 1 shows that the average standard deviation of logarithms is $s_o = 0.449$, and that the standard deviation of standard deviations is

$$s(s_o) = \frac{s_o}{\sqrt{2n_o}} = \frac{0.449}{\sqrt{50.8}} = 0.0626 \qquad (25)$$

where the average length of observation is taken to be $n_o = 25.4$ years.

The use of the average length of records for these stations instead of using the length of record of each station as a weight gives a more conservative test. It should be expected that, if the records vary around the mean length in a large range, the variation of the slope will be greater than if all records have been of the same average length.

Taking a 95% confidence level for the statistical test, with $t = 1.96$ for the standard normal distribution, then $t \cdot s(s_o) = 1.96 \times 0.0626 = 0.123$,

*The assumption of normal distribution of the slope is only an approximation for an adequate number of stations (say, more than 10)

M E D
12-1967

TABLE 1

Station No.	Name of Watershed	Estimate of the reliability -class-	Standard deviation of logarithms of peak* s	Length of observation n	n·s
1	Arroyo Secco	A	0.416	65	27.10
3	San Francisquito	A-B	0.371	26	9.65
4	San Lorenzo	A	0.460	30	13.80
5	Upper Saratoga	A	0.412	34	14.00
8	San Lorenzo Creek	A	0.500	20	10.00
9	Cayote	B	0.408	25	10.20
10	Uvas	A	0.389	27	10.45
15	Branciforte	A	0.402	17	6.84
16	Pescadero	B	0.503	15	7.55
17	Soquel	A	0.468	15	7.04
32	Matadero	B	0.560	15	8.41
34	Los Gatos	B	0.496	15	7.45
			5.385	304	132.49

$s_o = 0.449$ $\quad n_o = 25.4 \quad s_o = \frac{132.49}{304}$

$s_o = 0.435$

*Standard deviations of frequency curves are graphically estimated. The main weight has been given to the points of probabilities between 10% and 90% when drawing a straight line fit to frequency curves.

MICROFILMED
10-1967
PG 54 OF 73

so that the confidence limits are $s_1 = 0.449 + 0.123 = 0.572$ and $s_2 = 0.449-0.123 = 0.326$. From 12 stations in Table 1, 5% of 12 is only 0.6 or 1 station. Therefore, it should be expected that, on the average, only one station out of twelve should be outside the confidence limits 0.326 - 0.572. None of the s values in Table 1 is outside these limits. Therefore, one can conclude that the slopes of all 12 frequency curves are not significantly different from the constant value of s = 0.449.

Taking into account the second paragraph, that is using the sample size n as the weight, the average value of s is $s_o = 0.435$ (See Table 1). For $n_o = 25.4$ (average sample size), then,

$$s(s_o) = \frac{s_o}{\sqrt{2n_o}} = \frac{0.435}{\sqrt{50.8}} = 0.0605 .$$

For 95% confidence level and t = 1.96, the two confidence limits are

$$s_1 = 0.435 + 1.96 \times 0.0605 = 0.553$$

$$s_2 = 0.435 - 1.96 \times 0.0605 = 0.317 .$$

Only one value of s, that of Matadero Creek in Table 1 has s = 0.560 outside the confidence limits, which should be expected, because 5% of all s-values should be outside the limits and 5% is approximately one station.*

If the variance s^2 and n as weights are used, then the values in Table 1 give

$$\sum_1^{12} ns^2 = 58.31; \qquad s_o^2 = \sum_1^{12} ns^2 / \sum_1^{12} n = 58.31/304 = 0.192.$$

Then $s_o = 0.438$, which is approximately the same value as a weighted s.

*In a recent run by the District, it was found that the slope is related to precipitation, with a R^2-value of about 0.50-0.60. This would imply that the slope is a variable parameter.

MICROFILMED

Therefore, there is no evidence--in this case-- that the slopes of frequency curves are significantly different from a constant. This constant lies somewhere between 0.415 - 0.445. The value 0.430 or 0.435 seems the most likely average of the above stations.

In the study of regression equations for the s-parameter, the District has found that the standard deviation of logarithms of flood events correlates only with the mean annual precipitation, P. The values of the multiple correlation coefficient, R, were found to be 0.341 for flood peak discharges, 0.495 for 1-day and 0.504 for 2-day flood volumes, but precipitation is negatively correlated. This would imply that the smaller the mean annual precipitation, the larger is the slope of the frequency curve. In other words, the smaller the mean annual precipitation, the higher is the fluctuation of floods from year to year. This implied relationship from the correlations may be contested on several grounds, though it may be logical.

The correlation coefficient $R = 0.504$ (greatest of all three found) means that only 25% of variation of the slope (s) can be explained by the mean annual precipitation. It is difficult to use a prediction regression equation on the basis of regression analysis with an explained variance so low. A general rule in hydrology is that regression prediction equations with $R < 0.7$ (or $R^2 < 0.50$) should not be used.

The following is recommended to the District:

1. To use a constant value of the slope of flood frequency curves for the interim hydrology report, if it comes out to be a constant.

2. This constant should be determined separately for flood peaks and 1-day and 2-day flood volumes with greater weights for these latter two.

3. The mean of these three constant values should be used for frequency curves of peak, 1-day and 2-day volumes with greater weights for the latter two because they are more accurate than peaks.

MICROFILMED
12-1967

4. In determining the average values of s, the runoff data should be used selectively. It is noticed that the smaller the sample size, the larger are deviations between values of s and the average s-value, which should be expected. Using the lengths of record as weights in determining the average values of s_0 is recommended. This gives a lower value of s_0, which is close to the s-value for the Arroyo Secco, the stream with the longest record.*

5. For future revisions of the District's interim hydrology report, when more data will be available, the constancy of slope of frequency curves should be thoroughly investigated and tested. If the inclusion of various other parameters, except the mean annual precipitation, shows at least a value of $R = 0.70$, the prediction regression equation for s should be used. The use of isopleths may then be more justifiable than in the case of low values of R. By the time of that revision and the supplemental work (say, in about 5 years, or by 1972), there will be much more data on precipitation, especially on its intensities, duration, and areal coverage. It may be that some other precipitation parameters, besides the mean annual rainfall, may develop as significant independent variables in the multiple linear regression equation of the type of Eq. (19).

4. <u>Remarks on the use of step-wise method in selecting the order of significant independent variables</u>

It is known that the step-wise method in selecting the order of significant independent variables in multiple regression analysis has a built-in bias. Even for no relationship of a set of variables, the step-wise method would select that independent variable which by pure sampling chance shows the highest positive or negative partial correlation. Therefore, whenever several independent variables become significant, the F-test or

*This discussion is based on the assumption that s_0 would come out not to be significantly different from a constant.

MIC MED
12-1967
PG 57 OF 78

any other test of significance should be carried out for every independent variable to be included into the regression equation.

5. Potential new independent variables

In order to improve the prediction regression equations for flood events in Santa Clara County, it is desirable to include other independent variables besides the area A and the mean annual precipitation P. The brochure: "Research data assembly for small watershed floods " [11], lists a large number of potential independent variables. It is recommended to use as many of them as will be feasible to obtain future improvement of flood frequency information. This may not be practical for the forthcoming interim hydrology report. However, if q = Q/A is studied for the interim report, it is suggested that only those additional independent variable which are readily available or are easy to compute should be used besides A, P, S and Sh.

How the various additional independent variables may affect the prediction equations of flood frequency parameters, and which ones are likely to be significant, can be seen from several theses and reports on floods from small watersheds from Colorado State University, references [12] through [17]. Therefore, none of the potential additional independent variables are discussed here, seeing that enough information is contained in references [12] through [17].

6. Skewness coefficient of frequency curves

The estimate of skewness coefficient as the average regional value [18] is a proper approach when the flood frequency function requires three independent parameters (mean, standard deviation and skewness coefficient), as in the case of application of the Pearson Type III function with three parameters. For the log-normal distribution with two parameters (mean of

MIC LMED
12-1967
PG 58 OF 75

logarithms and standard deviation of logarithms), the skewness coefficient is a unique function of the coefficient of variation C_v of flood events (and not of their logarithms), and is changed every time C_v is changed. Therefore, in case of the use of log-normal distribution, the search for an average skewness coefficient valid for a region is not necessary and would conflict with the properties of the log-normal distribution.

As

$$1 + C_v^2 = e^{s_n^2} \tag{26a}$$

with s_n = standard deviation of logarithms, the skewness coefficient is dependent only on s_n. If the idea of a constant value of s_n is accepted for the interim hydrology report, then the skewness coefficient of log-normal distribution also becomes a constant. As some of the latest computations by the District show that s_n may not be a constant, but significantly related to the annual precipitation (with a negative power), the skewness coefficient then also changes from watershed to watershed. This fact will then negate the justification for using a constant C_s even for the Pearson TypeIII probability distribution.

7. Significant independent variables for prediction equations

In evaluating whether an independent variable in step-wise regression analysis should be included in the prediction equations of mean and standard deviation of logarithms, four statistics should be computed and may be used for each independent variable to be added eventually to previously selected independent variables by the step-wise procedure:

(a) F-statistics for F-test;

(b) σ_e - standard error of regression equation;

(c) R^2 - value (coefficient of determination); and

(d) P - the probability that the independent variable under investigation does not significantly contribute to the explained variance of the dependent variable.

The critical values of these four parameters should be selected in advance. For σ_e, the independent variable with the smallest value of σ_e should be the last to be included into the prediction equation. R^2 should be a minimum of 50%. P should not be greater than about 1% - 5%. Similarly, F-statistics value should be selected. All four criteria for the acceptance or rejection of independent variables will not be equal, and a compromise for the last independent variable to be included must be made by assessing the physical importance of the independent variables to be excluded from the prediction equation.

8. Effective number of independent stations

If the station-year method is used for any flood analysis, the fact is that the series of annual flood values may be considered as serially uncorrelated. However, as annual flood series are correlated among stations, the effective number of independent stations can be computed by

$$N_e = \frac{N}{1 + \bar{r}\,(N-1)} \tag{26b}$$

where N = number of stations for which the flood events are pulled together, $\bar{r}$ = average correlation coefficient of all correlation coefficients--altogether N(N-1)/2 different values--between annual flood series of N stations, N_e = effective number of stations. The pulling together of N dependent stations, with an average number of years n_o gives an equivalent of independent station-years

$$n_e = n_o N_e = \frac{n_o N}{1+\bar{r}\,(N-1)} \tag{26c}$$

MICROFILMED
17-1967
PG 60 OF 73

where n_e can be used in statistical tests and drawing-up of confidence limits instead of n in previous equations.

This procedure may be used,--with n_e = effective number of independent station years,--in the case the District decides to use the slope (s) as a constant for all watersheds, at least as one solution of the prediction problem, with N = number of watersheds from which the constant s is determined, $\bar{r}$ = their average correlation coefficient of flood series, and n_o = average length of time series, as was shown in Table 1. Then n_o in eq. (25) should be replaced by n_e of eq. (26c). It will then enable a drawing of confidence limits with s_1 and s_2 (confidence limit slopes) being constant for a given confidence probability level.

CHAPTER IX

REGIONAL DISTRIBUTION OF REGRESSION CONSTANTS (ISOPLETHS)

1. District's procedure

The basic hypotheses which underlie the concept of regression constants and isopleths are:

(a) The prediction regression equations for mean and standard deviation of logarithms of flood events do not include numerically many hydrologic factors, such as soils, vegetable cover, elevation, climatic factors, antecedent moisture conditions, various geological properties, several geometric characteristics of watersheds and similar. Therefore, the remaining unexplained variance of m and s was contained in the regression constant.

(b) The regression constant varies from watershed to watersheds.

(c) Constants can be determined for each drainage basin by using the general equation derived from the prediction equations like those of Eqs. (21) through (23) which are slightly modified and from the basin statistics (computed mean and standard deviation of logarithms).

The prediction regression equations are modified to the form:

$$Q_{mp} = 0.01\ C_p\ A^{0.8}\ P^{1.5} \tag{27}$$

$$s_p = 0.01\ K_p\ P^{-0.214} \tag{28}$$

$$Q_{m2} = 0.001\ C_2\ A\ P^{1.8} \tag{29}$$

$$s_2 = 0.01\ K_2\ P^{-0.25} \tag{30}$$

where Q_{mp} and s_p relate to flood peaks, Q_{m2} and s_2 to 2-day flood volume, and C_p, K_p, C_2 and K_2 are corresponding regression constants. The values Q_{mp}, s_p, Q_{m2} and s_2 represent the statistical parameters computed for each

M E D
12-1967
PG 62 OF 13

individual drainage basin. From Eq. (27) through (30) and similar equations for 1-day flood volumes, the regression constants are determined for each individual watershed.

The regression constants were plotted in centers of their respective drainage basins. Then a smoothing technique was used to take into account the errors in data and in the estimates of the above parameters. Isolines, called isopleths, were then drawn for each parameter m or s, and for each of the three flood characteristics (peak, 1-day and 2-day volume).

The isopleths were drawn so that the basin's regression constant was distributed with respect to basin area; i.e. by summarizing the areas between isopleths, the basin constant could be retrieved.

The procedure for determining the peak discharge for a design frequency within the study area is as follows:

(1) Solve for the standard deviation using the equation

$$S_p = 0.01\ K_p/P^{0.214} \tag{31}$$

and for the mean using the equation

$$Q_{mp} = 0.01\ C_p\ A^{0.8}\ P^{1.5} \tag{32}$$

The precipitation is obtained from isohyetal maps based on long records, the drainage area from topographic maps, and the residual constants from the isopleth maps.

(2) With the knowledge of the standard deviation and the mean, the frequency curve can be computed. An adjustment is made for the effects of using a small sample drawn from a normal parent population. This adjustment reflect the weighting of probabilities that differ from that of maximum likelihood (see Beard JGR-V65-July 1960).

MICR... ...MED
12-1967
PG 63 OF 70

(3) These results are plotted on log-normal probability paper and flows for any frequency can be picked from the curve.

2. General criticism of use of regression constants

Though the use of regression constants is extensive in hydrologic practice, it is a very sensitive and controversial technique. The assumption underlying the technique is that the regression equation takes into account some independent variables but not all. Therefore, the differences between the predicted values by the regression equation and the observed values, as the deviations from the regression equations, may be represented after some smoothing as isolines over a region. This approach assumes that there is no sampling variation in parameters and no errors in data, or that they are small, or that they may be eliminated by the smoothing procedure. Therefore, it is not surprising that isopleths of these deviations, called in the District's work the regression constants, have very erratic and often unexplainable patterns.

The deviations from regression equations contain the following components: (1) Error component coming from errors in measurement and computation of floods, which can go from 10% for the best observed floods to 25-50% for floods estimated by unreliable rating curves and for other errors in flood values; (2) Sampling error component produced by the sampling errors of parameters (mean and standard deviation), which are very large for short records. As is shown in the discussion of sampling variation of the slope of frequency curves, it is inevitable that both parameters of frequency curves which are used in the District's analysis contain substantial sampling errors; (3) The component resulting from the neglect of several factors in regression analysis. This may account for a part of variation of the mean and standard deviation of flood frequency curves.

MICROFILMED
12-1967

The concept of a regression constant is justified only by the third factor of unaccountable independent variables. This third factor usually accounts only for a relatively limited portion of the total unexplained variation of the mean and standard deviation.

The use of isopleths of smoothed regression constants means a smoothing of errors in flood determination, and of sampling errors of frequency curve parameters. The use of smoothed regression constants and the corresponding isopleths means that the next drainage basin should have approximately the same unaccountable factors of regression analysis, with nearly the same effect as the nearest gaged drainage basin. This is unlikely to be the case. The method of regression constants and their isopleths should be very carefully used. When it can be assumed that the logarithms of flood values have small errors and small sampling fluctuations, this method of isopleths may be used. The factors unaccounted for by the regression analysis and related to precipitation and basin characteristics should change slowly from one drainage basin to another. When the change in unsmoothed regression constants is very gradual, say from one side of the mountain to the next, or from the valley to the top of mountains, one would be tempted to recommend the use of this method of regression constants and isopleths.

3. Use of regression constants in the District's work

If the prediction equations (21) through (23) are used, with 82.5 - 87.0% of variance of Q explained by these equations, it may not be necessary to use the regression constants and isopleths at all. In this case, it is sufficient to estimate the values of Q by Eqs. (21) through (23) from the drainage basin area and its mean annual precipitation.

MICROFILMED
12-1967
PG 65 OF 78

Whether the correction should be made for the difference in accuracy between the estimation of the mean of logarithms of floods for various drainage basins by the method of moments and by the method of maximum likelihood, as suggested by L. Beard and quoted above, depends on the skewness coefficient. The smaller the skewness coefficient, the less important becomes this difference [20], for $s_n = 0.435$, then

$$C_v = (e^{s_n^2} - 1)^{\frac{1}{2}} = 0.460$$

The skewness coefficient is then

$$C_s = 3\,C_v + C_v^3 = 1.380 + 0.096 = 1.476 \simeq 1.5$$

Because of relatively high values of C_s, the corrections for the better estimates by maximum likelihood is justified for the District's study.

If the prediction equations for the standard deviation of logarithms are used with such low R^2-values (maximum is 0.25), then more than 75% of its variation is unexplained. It may then be attractive to use regression constants and isopleths for the s-value, because one would expect to explain such a large unaccounted variation in s by this means. However, the above test of variation of s, and the conclusion that s is not significantly different from a constant, makes the use of regression constants and isopleths unwarranted.*

If q = Q/A is used as the dependent variable and as the central value parameter of frequency distribution, it is always significantly different from a constant. Its regression analysis may show an R-value somewhere around 0.70. In this case, the unaccounted variation by other drainage basin factors may warrant a use of regression constant and isopleths.

*An attractive approach may be the use of s as a constant, for one flood prediction solution, and the prediction equation of s (with R^2 = 0.50-0.60 at least) for the second flood solution in determining the slope of frequency curves of ungaged watersheds.

MICROFILMED
12-1967

CHAPTER X

FLOOD HYDROGRAPHS

1. District's procedure

In the case where routing studies are involved, a design hydrograph, as well as the design peak, is required. To accomplish this, a model hydrograph is developed for the area. This hydrgraph is based on the "Standard Project Storm" for the area and the unit hydrograph correspondning to the area. The unit hydrograph is developed using the Clark method, with the Clark constants being defined from the results of using an optimization technique for studying historical floods of similar adjacent basins. This model hydrograph is then adjusted to correspond with the design frequency for the peak, 1 hour, 6 hours, and so on. The resulting balanced hydrograph is a design hydrograph that corresponds to one frequency for all durations.

As a check on larger basins, a unit hydrograph (developed as noted above) is utilized to calculate a flood hydrograph for a synthetic rain flood of amounts and distribution which have historically created major floods. This flood hydrograph characteristic should exceed those obtained from the 100-year balanced hydrograph.

2. Discussion of balanced hydrograph

The 100-year return flood peak and the 100-year return flood volumes may be assumed to occur at the same time*. This is a conservative approach in the sense that the balanced hydrograph composed of simultaneous occurrence of 100-year peak and 100-year volume may require larger reservoir flood control capacity. A study of the type of flood hydrographs which occur on streams in Santa Clara County shows that there may be a large correlation of the peak discharge and flood volumes. However, that correlation

*If the correlation coefficients between peaks and 1-day, and peaks and 2-day volumes were of the order 0.90-0.95, this approach would not depart significantly from the reality.

is not likely to be in the range of 0.90-0.95. Therefore, there is always a random element which makes the coincidence of 100-year peak and 100-year volume less probable. The District should use simultaneously 100-year return periods for both peaks and volumes as a built-in safety factor.*

The use of unit hydrographs and standard project storms for the area is a classical approach used by the U. S. Army Corps of Engineers. The district should continue to use them, at least for the forthcoming interim hydrology report. One objection to this application is the types of storms and flood hydrographs in Santa Clara County.

By scanning through many observed flood hydrographs in the County, it is easy to notice two types of hydrographs:

(a) Well developed one-peak hydrographs, for which the application of simple standard project storm and of unit hydrograph can be easily justified; and

(b) Flat and long-duration multi-peak hydrographs, for which it would be difficult to adjust a simple standard project storm and use the unit hydrograph.

For purposes of design of channels by the flood routing approach, the use of one-peak hydrograph and the procedure outlined by the District seems to be the best approach to follow in the forthcoming interim hydrology report.

3. Other possible approaches

Some other approaches for the prediction of flood hydrographs may be used once the peak discharge and flood volumes are known. They may be applied in the future updating and improvement of the interim report by an extensive investigation of types of storms and flood hydrographs. A proper approach may be a prediction of the total volume of flood events,

*It is suggested that the correlation coefficients between peaks and 1-day (or 2-day) flood volumes be computed to check the likelihood of coincidence.

MICROFILMED
12-1987

of the rising time for its rising limb, and of the type of recession curves. For this approach the references [12] through [17] may be useful.

For these alternate approaches, the use of a relationship of parameters for flood peak, 1-day flood volume and 2-day flood volume (three points for times 0, 1-day and 2-days enable the drawing of a curve) can be very beneficial. It can produce flood volumes of 6-hours, 12-hours and so forth. These flood characteristics would then enable one to assemble synthetic flood hydrographs of 100-year return period, or any other return period. For the future improved prediction of flood hydrographs, two types of design hydrographs of 100-year return period (or any other return period) may be necessary and useful:

(a) One-peak hydrograph, designed with the flood peak of 100-year return period, but following the unit hydrograph method and an adjusted standard project storm in order to produce the predicted flood peak, but neither required to match the predicted 1-day and 2-day flood volumes of 100-year return period nor any other time interval flood volume derived from these basic flood characteristics (peak, 1-day and 2-day volumes).

(b) Use of the predicted 1-day and 2-day flood volumes of 100-year return period (or any other return period) as the basis for reconstructing a balanced (synthetic) hydrograph, without matching the predicted flood peak of 100-year return period, and without using the standard project storm (either adjusted or unadjusted) and the unit hydrograph. The use of other hydrograph properties (total flood volume, rise-to-peak time, recession curve properties and similar) would produce a flat and long-duration hydrograph. This will give an extreme of the type of flood hydrographs on streams of Santa Clara County, the other of which will be the one-peak hydrograph, as described under (a).

LMED
1987
64 OF 78

The one-peak hydrograph of 100-year return period may be predominately (but not exclusively) used for the flood routing along channels and for the design of channel capacities. The flat and long-duration hydrograph of 100-year return period may be predominately used (but not excessively) for the design of new flood control reservoirs, or for determining the effect of existing reservoirs on flood attenuation.

The suggestions outlined above under the two alternatives (a) and (b) may seem contradictory with the previous recommendation to use simultaneously the 100-year flood peak and 100-year 1-day and 2-day flood volumes in the balanced hydrograph for design purposes. However, the latter two recommendations under (a) and (b) will produce a more realistic design, but the built-in-safety will not be included. If the District uses the 100-year upper confidence limit flood (say of 75% or 80% confidence level) to see whether the designed structure would perform at an adequate safety, the built-in-safety of the simultaneity of 100-year peak and 100-year 1-day and 2-day volumes is not necessary.

There is not enough knowledge about the precipitation structure of storms which produce the floods of rare occurrence in the area of Santa Clara County. It seems that on the general frontal storms the existence of individual storm-cells is the basic pattern of large flood genesis.

CHAPTER XI

EFFECTS OF URBANIZATION

1. The continuous process of urbanization

Experience shows that the urbanization is a continuously growing process in Santa Clara County [19]. It goes on both at the County floor and in the surrounding hills. With time there will be only three types of watershed areas in the County:

(a) Mountains, mainly forested and unpopulated, or sparsely populated watersheds;

(b) Hilly populated areas, with control drainage by terracing, slope protection, improvement of steep gully channels; and

(c) County floor areas intensively developed and urbanized. Naturally, all transitions between these three types will continue to exist.

As a consequence of intensive urbanization, the floods will change appreciably even in the time interval of one generation.

Theoretically speaking the floods should increase with the degree of urbanization both in its peaks and volumes. The peaks will increase because:

(a) The infiltration decreases and also the general losses decrease so that the effective precipitation (surface runoff) increases;

(b) Less detention is available because of regular, well defined and well drained streets and other paved areas, large roof areas, etc. so that the attenuation of surface runoff decreases; and

(c) The urbanized paved and roofed areas, including gutters, storm drains and regulated channels decrease the general roughness coefficients in watershed flow. This may increase the peakedness of floods.

C. ED
12 7

The above factors but primarily the decrease in infiltration and losses lead to an increase of flood volumes.

2. District's procedure to determine the effect of urbanization on floods

The District's procedure takes into account the encroaching urbanization by transforming the flood frequency curves, representative of mountain and foothill watersheds, to those of mixed areas.

The general approach is to determine the discharge-frequency relationship for the total area above each point or reach in which a design flow is required. Similar to the case where there is reservoir control above the point of interest, the urbanization approach requires the development of a design hydrograph for the total area. This is accomplished by using the model hydrograph concept and developing hydrographs for sub-areas using the Clark unit hydrograph method and the Corps of Engineers standard project storm. These sub-area hydrographs are individually routed and combined to make up the total area model hydrograph. The urbanization effect is accounted for by modifying the loss equation used in obtaining run-off from the standard project storm. This equation, with losses equal a constant times precipitation to the x power, will give the losses to be expected from a large storm (one percent probability flood, or Standard Project Flood), with normal type cover. In the District's analysis the computed ratio of effective impervious area is based on the State of California ITTE Report and the Santa Clara County Planning Department projected development. The revised losses are then computed by the equation:

$$\text{Losses (Revised)} = \text{Losses (Normal cover)}(1 - r)$$

with r = the ratio of effective impervious area to total area. These revised losses will be smaller, resulting in an increased runoff from areas affected by urbanization.

The percent of increase of the model hydrograph, due to urbanization over that for normal land cover is used to adjust the design hydrograph. This design hydrograph is developed by using the discharge frequency curves (peak, 1-day and 2-day values for the one percent probability flood), and the model hydrograph. This technique is used for both the foothills and the valley floor. In a recent study when the described method was applied, the computation showed a 20 percent increase in the peak discharge for a highly urbanized subarea (approximately 1.5 square miles).

Alternatively, to describe urbanization effects for low lying areas through which a leveed flood control channel is constructed, a unit inflow per mile of leveed channel was applied. This inflow is handled as an inverse channel loss in the routing process and is representative of a constant local inflow resulting from temporarily flooded streets and storm sewers contributing through surcharged inlets into the channel.

3. Comments on the above District's procedure

The use by the District of the model hydrograph concept and development of hydrographs for sub-areas of a watershed is the proper approach. Sub-area hydrographs and standard project storms take into account any particular urbanization degree of any sub-area in the County. The individually routed sub-area hydrographs, combined to make up the total area model hydrograph, produce a total revised flood hydrograph which takes into account the urbanization of sub-areas.

The determination of a new loss equation by modifying the preurbanization loss equation is a key step used in the District's methodology to take the urbanization into account. The loss equation is of the form

$$L = K P^{x} \tag{33}$$

with L = losses; P = precipitation of a large storm; K = a constant; and

CROPPED
12-1.
PG 22 OF 11

x = power of the relationship. This gives a simple straight line regression relationship between log L and log P. Equation (33) should be considered only as a first approximation in evaluating the loss equation in floods for watersheds of Santa Clara County, in general, and of the loss equation considering the process of urbanization of the County, in particular. It should be expected that antecedent moisture conditions should, especially, be considered in the loss equation, as well as other parameters besides the total storm precipitation. It is a sound procedure to assume that the search for the best loss equation should begin with a rough first approximation and then should be improved by the principle of successive approximations.

Because of the limited time before the pending interim hydrology flood prediction report of the District is finished, it is recommended that the District uses the developed loss equation, Eq. (33), bearing in mind that it is a first approximation. However, for future revisions and improvements of the interim report, it is suggested that a more elaborate loss equation be developed by including other important drainage basin factors in floods, which become significant variables in regression equations.

The loss equation is a very important part and quite likely the most predominant factor in determining the flood (1-day and 2-day) volumes. However, for peak discharges, besides the loss equation, the change of surface storage conditions and the change in resistance factors in the overland flow and in conveyance structures of urbanized areas may not be negligible factors. Such variables are not taken into account by the simple loss equation in which log L is linearly dependent on log P as only independent variable.

In the absence of better techniques for measuring the urbanization effects on floods, the District's approach of using the ratio C_r of effective

impervious area to the total area (called RTIMP in the District's methodology) is a reasonable approach as the first approximation. It is endorsed, therefore, for the interim hydrology report. For the future program of improvement (a 5-year or similar period) and for up-dating of the interim report, it is recommended to introduce other factors of urbanization (especially storage) and resistance factors in addition to the percentage of impervious area.

The following equation is used for revised losses, used by the District, is

$$L_r = L_n (1 - r) \tag{34}$$

where L_r = revised losses, L_n = losses under natural watershed cover, and r = ratio of impervious to total area.

A simple assumption underlies this equation, namely that all types of water losses from impervious areas are negligible. Besides the zero infiltration (or so), the evaporation is very small during a storm from these impervious areas. Also, the eventual flow from the impervious areas over the pervious area would not increase the infiltration or evapotranspiration losses over the pervious areas. This basic assumption is correct for a first approximation taking into account the various approximations already made regarding other aspects of urbanization. However, for future revisions of the interim report, this simple approach should be checked and eventually improved.

It may be beneficial for the District to establish a 5-year program for the investigation of effects of urbanization on floods in the County. The greater the urbanization, the more important is a proper and accurate estimate of flood characteristics. The more extensive urbanization, the larger are the investments per unit area, and the greater the effect

of urbanization on floods. A pilot project consisting of small adjacent areas, one under natural conditions and the other highly urbanized, or 2-3 urbanized small areas with various degrees and types of urbanization, observed for a couple of years would give very valuable data and the corresponding prediction equations for the effects of urbanization. It is suggested that such a pilot project should be seriously considered under the rainfall regime and with the type of urbanized areas that occur in Santa Clara County.

The use of modified loss equation to obtain a modified design hydrograph as is done by the District is a good approach. It should not be surprising that the modified flood hydrographs of highly urbanized areas may have computed peaks of 20 percent or more than for the pre-urbanization conditons. Also the handling of local inflows along the leveled flood control channels as an inverse channel loss during the flood routing is a correct approach.

September 6, 1967
Fort Collins, Colorado

Vujica Yevjevich
Darryl B. Simons

MICROFILMED
PG 76 OF 78

REFERENCES

1. Hydrologic Data Index, Santa Clara County, California, prepared by Carroll E. Bradberry and Associates, Consulting Engineers, Los Altos, California, 1963 Revision.

2. Hydrologic Atlas, Santa Clara County, California, prepared by Carroll E. Bradberry and Associates, Consulting Engineers, Los Altos, California, 1964 Revision.

3. Synthesized Streamflow Data, Santa Clara County, California, prepared by Carroll E. Bradberry and Associates; Consulting Engineers, Los Altos, California, 1964.

4. Summary Report, Hydrologic Data Program, Phase III - Streamflow Synthesis, Santa Clara County, prepared by Carroll E. Bradberry and Associates, Consulting Engineers, Los Altos, California, 1964.

5. Santa Clara County Flood Control and Water Conservation District Hydrology Study - Las Animas Creek, memorandum prepared by Water Resources Engineers, Inc., Consultants, Lafayette, California, December 27, 1965.

6. Santa Clara County Flood Control and Water District: Reconnaissance Report on Floods of January 21-24, 1967, Issued January 1967, San Jose, California.

7. U. S. Corps of Engineers, San Francisco: San Francisquito Creek, Flood Control and Allied Purposes, January 1961, San Francisco.

8. U. S. Department of Agriculture and California Agricultural Experiment Station: Soil Survey, Santa Clara Area, California, Series 1941, No. 17.

9. U. S. Army Corps of Engineers, Sacramento, California: Sacramento-San Joaquin Valley, California, Standard Project Rain-Flood Criteria, April 1957, Revised September 1958.

10. Dalrymple, Tate: Flood-Frequency Analyses, Manual of Hydrology: Part 3. Flood-Flow Techniques, Methods and Practices of the Geological Survey, U. S. G. S. Water-Supply Paper 1543-A, 1960.

11. Laurenson, E. M., Schultz, E. F. and Yevjevich, Vujica, Research Data Assembly for Small Watershed Floods: Report by Engineering Research Center, Colorado State University, Fort Collins, Colorado, September 1967.

12. Siddiqi, M. K., Rainfall-Runoff Volume Relationship for Small Watershed Floods, M. S. Thesis, Colorado State University, Fort Collins, Colorado, August 1966.

MICROFILMED
12-1967
PG 77 OF 78

13. Voytik, Andrew, Runoff Predictions from Arid Regions, M. S. Thesis, Colorado State University, Fort Collins, Colorado, June 1967.

14. Om Kar, Songthara, Hydrograph Rise Times, M. S. Thesis, Colorado State University, Fort Collins, Colorado, June 1967.

15. Ho, Yu Bing, Evaluation of Runoff Characteristics From Hydrograph Recessions, M. S. Thesis, Colorado State University, Fort Collins, June 1967.

16. Bell, F. C., Extreme Rainfall of Short Duration, manuscript submitted for publication in the Journal of the Hydraulics Division of the American Society of Civil Engineers, May 1967.

17. Bell, F. C., Estimating Design Floods From Extreme Rainfall (With Special Reference to Small Watersheds in Western U.S.A.), prepared for the Bureau of Land Management, Colorado State University, Fort Collins, Colorado June 1967.

18. Beard, Leo R., Statistical Methods in Hydrology, revised edition published under Civil Works Investigations Project CW-151 by the U. S. Army Engineer District, Corps of Engineers, Sacramento, California, January 1962.

19. Harris and Rantz, Effect of Urban Growth on Streamflow Regiment of Permanente Creek Santa Clara County California, Geological Survey Water-Supply Paper 1591-B.

20. Thom, H.C.S., A Note on the Gamma Distribution. Weather Bureau, Monthly Weather Review, 86:117-122, 1958.